A Survey of Quantum Theory Inspired Approaches to Information Retrieval

A Survey of Quantum Theory Inspired Approaches to Information Retrieval*†

SAGAR UPRETY, The Open University, UK
DIMITRIS GKOUMAS, The Open University, UK
DAWEI SONG, Beijing Institute of Technology, China & The Open University, UK

Since 2004, researchers have been using the mathematical framework of Quantum Theory (QT) in Information Retrieval (IR). QT offers a generalized probability and logic framework. Such a framework has been shown capable of unifying the representation, ranking and user cognitive aspects of IR, and helpful in developing more dynamic, adaptive and context-aware IR systems. Although Quantum-inspired IR is still a growing area, a wide array of work in different aspects of IR has been done and produced promising results. This paper presents a survey of the research done in this area, aiming to show the landscape of the field and draw a road-map of future directions.

CCS Concepts: • **Information systems → Information retrieval**; **Document representation**; **Retrieval models and ranking**; **Language models**; **Multimedia and multimodal retrieval**; **Users and interactive retrieval**;

Additional Key Words and Phrases: Information Retrieval, Quantum Theory, Quantum-inspired models

ACM Reference Format:
Sagar Uprety, Dimitris Gkoumas, and Dawei Song. 2020. A Survey of Quantum Theory Inspired Approaches to Information Retrieval. *ACM Comput. Surv.* , (2020), 37 pages.

1 INTRODUCTION

Information Retrieval (IR) is the process of finding information that is relevant to the need of a user. The last two decades have completely changed how humans consume and interact with information. This change has been driven by the advances in web search engines, the ease of access to the Internet and the explosion of information available online. Information pertaining to a variety of needs is available - from lecture slides to news articles to descriptions and reviews of items, and so on. It becomes imperative that the IR systems continually improve to accommodate such information needs, which have been growing both qualitatively (in terms of complexity) and quantitatively. Essentially, the task of IR systems can be reduced to two aspects. One is how to efficiently and effectively represent and rank the variety of unstructured information being created at each instant. This involves tasks like indexing and improved understanding of the content through advanced representation methods, as well as ranking of the information items based on the representation. For example, representation of textual information can be improved with better understanding of natural language. The second is how to make an IR system better understand user's complex information need and information

*This work is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 721321, and Natural Science Foundation of China (Grant No U1636203). Corresponding author: Dawei Song.
†This manuscript has been accepted for publication at ACM Computing Surveys on May 20, 2020

Authors' addresses: Sagar Uprety, The Open University, UK, sagar.uprety@open.ac.uk; Dimitris Gkoumas, The Open University, Milton Keynes, UK, dimitris.gkoumas@open.ac.uk; Dawei Song, Beijing Institute of Technology, China & The Open University, UK, dawei.song2010@gmail.com.

seeking behaviour. This involves understanding user's search context, search task and intent, and ability to measure task completion and user satisfaction through user interactions.

IR researchers have been investigating different approaches to improve IR systems from both the system point of view (representation and ranking) and the user point of view. Various areas in IR, for example, Neural IR [Mitra and Craswell 2018; Onal et al. 2018], Interactive IR [Borlund 2013; Ruthven 2009], Cognitive IR [Ingwersen 1996; Järvelin and Ingwersen 2012; Sutcliffe and Ennis 1998], and Dynamic IR [Sloan and Wang 2015; Yang et al. 2016b], have been developed. Quantum-inspired IR (QIR) is one such area, where the mathematical framework of Quantum Theory (QT) is utilized to develop representation and user models in IR that are expected to better align with human cognitive information processing. It is different from the field of Quantum Computing in that it does not involve computations based on physical quantum states.

The benefits of using QT in IR are many-folds. It offers a new way of representing events and computing probabilities of events. Instead of the set-theoretic method of representing events as subsets of a larger sample space, QT represents events as subspaces of an abstract, complex vector space (called Hilbert space) [Busemeyer and Bruza 2012]. Moreover, the same event can have multiple representations in multiple basis of the Hilbert space. This method of representation can help in the abstraction and contextualization of information objects like documents and queries [van Rijsbergen 2004]. For example, if a set of basis vectors correspond to a set of documents, another set of basis vectors in the same Hilbert space can represent the same set of documents in a different context. Hence a query (as an event) will be represented by these different basis depending upon the context of retrieval. The Hilbert space representation of events also leads to a generalized method of calculating probabilities (Born rule) [Born 1926], by taking into account interference between events. This can model a user's decisions under ambiguity better than traditional probability models [Busemeyer and Bruza 2012]. Such a representation method can inherently model incompatible variables - those where measurement on one variable affects the outcome of the other. For two such incompatible variables A and B, measuring A would alter the state of the system, so that the subsequent measurement of B would be different than if it was measured alone or before A. Thus these two variables cannot be measured simultaneously or jointly, and different orders of measurement would lead to different outcomes. Traditional probability theory assumes that for any pair of events, $p(A, B) = p(B, A)$, which would be incorrect for incompatible variables. The cognitive phenomenon of order effect is generally considered to be a consequence of incompatibility in measuring human decisions [Busemeyer and Bruza 2012]. There has been a lot of research in recent years, which shows the presence of Order Effects in relevance judgment of documents (detailed in Section 3).

Correspondingly, the application of QT to IR can be broadly divided into two subareas: (1) Representation and Ranking, and (2) User Interaction. Figure 1 shows a sketch of the overlap between traditional IR and Quantum-inspired IR, and their underlying components. We show traditional IR in terms of the two sub-areas, which overlap because user interactions like relevance feedback are often used in re-ranking tasks. In this sense, QIR overlaps with traditional IR as it is also divided into these two sub-areas. The difference comes in the tools used by QIR. It utilises the mathematical framework of QT including complex Hilbert space models for representation learning and quantum probability rules to model cognitive interference in document ranking.

QIR borrows heavily from concepts, models and techniques developed in the field of Quantum Cognition, especially in the modelling and incorporating the user interactions in IR. QT has been successfully applied to model and predict irrational human decision making and explain cognitive biases in human judgments [Busemeyer et al. 2011; Pothos and Busemeyer 2011, 2009; Pothos et al. 2013; Trueblood and Busemeyer 2011; Wang and Busemeyer 2013]. The emerging field of Quantum Cognition [Busemeyer and Bruza 2012] studies such quantum-like phenomena in cognitive and

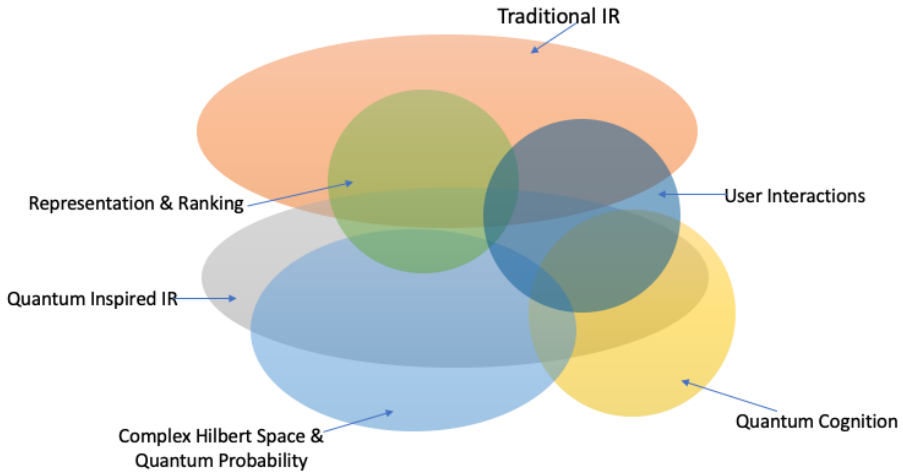


Fig. 1. Brief overview of quantum-inspired IR

decision sciences. There is already a growing community of researchers, under the umbrella of Quantum Interaction (see http://www.quantuminteraction.org/home), who are applying QT to various disciplines such as Biology, Cognition, Economics, Natural Language Processing and Information Retrieval. In 2017, a major project which seeks to investigate a Quantum Theoretical approach to IR (QUARTZ - see http://www.quartz-itn.eu/) has started, under the Marie Skłodowska Curie Actions scheme of the European Union's Horizon 2020 programme, with 7 participating universities all over Europe and several external partners around the world.

QIR is a growing multi-disciplinary area and has been attracting an increasing attention of researchers in IR. Especially, the recent several years have witnessed a large number of models and applications of QIR which have shown good results and a great potential. However, the field lacks a comprehensive review of the literature, and the individual works are largely segmented. This is why a survey paper is urgently needed. It is important and timely to review the literature systematically to provide a clear picture of the landscape and a road-map for the future. Although this is not the first work to accumulate findings in QIR. Around ten years ago, a position paper [Song et al. 2010], organised QIR into three themes: frameworks, spaces, and interference. However, it was more on a conceptual level and the field of QIR was in its infancy. After a decade of development since then, the landscape of QIR research has significantly changed. A large number of more comprehensive and larger scale QIR approaches have been developed covering different aspects of IR, and have achieved remarkable experimental results.

The next section introduces the basic concepts and notations of QT, to enable readers to understand their usage in IR. Section 3 reviews the literature of Quantum-inspired IR, followed by a discussion on its shortcomings and benefits in Section 4. Section 5 concludes the paper by discussing future work directions in Quantum-inspired IR.

2 QUANTUM THEORY PRELIMINARIES

Quantum Theory is also regarded as a theory for calculating probabilities [Pitowsky 2006], which was developed in the first half of the twentieth century to explain the counter-intuitive probabilistic outcomes of experiments on microscopic particles. These results could not be explained using standard probabilistic models. Quantum Mechanics was later axiomatically organized by John von Neumann [von Neumann 1955], thus enabling it to be used as an abstract mathematical framework even outside of Physics. The fundamental difference between the classical and quantum

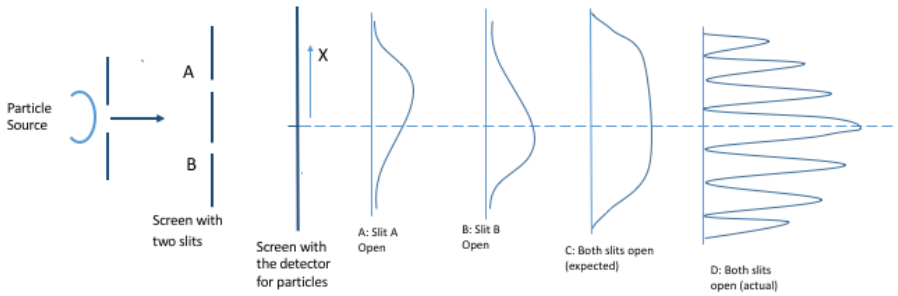


Fig. 2. Double slit experiment setup

probabilities lies in the representation of events. In the classical probability theory, events are represented as subsets of a sample space. In the quantum probability theory, events are represented as subspaces of an abstract vector space. As such, the quantum probability theory is a generalization of the classical probability theory, and can be useful in calculating the probabilities of events which cannot be represented in a set-theoretic formalism due to their inherent structure. The use of QT for applications beyond Physics was first suggested by Niels Bohr [Bohr 1937] (pg. 294-295, 297), one of the founding fathers of Quantum Mechanics. He mentioned the existence of complementary variables in Psychology as similar to the incompatible properties of quantum systems. As we will discuss in this section, QT provides a method to model incompatible variables naturally. In the sub-sections to follow, a brief description of the need for the quantum probabilistic framework is provided and the formal concepts underlying QT are discussed.

2.1 The Double Slit Experiment

The earliest experiment on microscopic particles which puzzled physicists was the Double Slit Experiment. Consider Figure 2, in which microscopic particles, say electrons, are fired from a source to a screen consisting of two slits. On the right of this screen is another screen made up of detectors, which can detect the arrival of a particle as a function of its distance x from the center of this screen. By measuring the mean number of pulses, one can measure the probability of the electron reaching the detector screen as a function of x. When only one slit is open, the probability distribution obtained looks like the one in Figure 2(A) and 2(B) for slits A and B respectively. Now, according to the classical probability theory, opening both slits at the same time should lead to the sum of the probability distributions as shown in Figure 2(C). However, it was actually seen that on opening both slits, one obtains a distribution of electrons as 2(D). It is a complicated curve having several maxima and minima, indicating that there are some locations on the detector screen that electrons never register. This distribution is the same as obtained in the case of interference of waves. The interference pattern is produced by adding the amplitudes of two waves and the squaring the sum to get the intensity. Hence the data of curve 2(D) can be explained by assuming that the electrons behave like waves when traveling from the source through the slits to the detector screen. In doing so, it is as if a single electron goes through both the slits at the same time - a fundamental quantum property called *superposition*. A complex number called the probability amplitude is ascribed to the electron corresponding to the two possible paths. Let ϕ_a be the probability amplitude for the path $S \rightarrow A \rightarrow X$ and ϕ_b be the probability amplitude of the electron for the path $S \rightarrow B \rightarrow X$, when the slits A and B are opened respectively, S being the source.

The amplitudes differ because of the difference in the complex phase for the two paths taken. The probabilities are calculated, according to the Born rule [Born 1926], as the square of the amplitudes. Thus the probability of detecting an electron at a position X from the center of the detector screen, when only slit A is open, is $p(A) = |\phi_a| * |\phi_a^\dagger| = |\phi_a|^2$

(where $\phi_a^\dagger$ is the complex conjugate of ϕ_a). When both slits are open, the probabilities are calculated by following the Law of Total Amplitude [Feynman et al. 2011]. The probability amplitudes for the two paths are added up and then the probability is calculated by taking the square of the sum:

$$\begin{aligned} p(X) &= |\phi_a + \phi_b|^2 \\ &= |\phi_a|^2 + |\phi_b|^2 + 2|\phi_a| * |\phi_b| \\ &= p(A) + p(B) + 2\sqrt{p(A)}\sqrt{p(B)}\cos(\theta) \end{aligned} \tag{1}$$

where θ is the phase difference between the two paths. The negative values of $\cos\theta$ are responsible for the minima obtained in the curve 2(D). For $\theta = \frac{\pi}{2}$, we get the classical probabilities as a special case.

Thus we see that the origin of quantum probabilities lies in the Law of Total Amplitude and the Born rule. When a quantum entity can take one or more paths, it takes all of them at the same time, and the quantum entity is said to be in a *superposition state* of all possible paths. These paths influence each other, in a phenomenon called *quantum interference*, which gives rise to the extra terms in the calculation of probabilities.

It should be noted that when we say quantum probabilities, the concept of probability remains the same as classical probabilities. To paraphrase Richard Feynman, "If the probability of a certain outcome of an experiment is p, then if the experiment is repeated many times one expects that the fraction of those which give the desired outcome is p. What changes in QT is only the method of calculating probabilities" [Feynman et al. 2011].

2.2 The Axioms of Quantum Theory

2.2.1 Representation of Events. QT provides a new method of assigning probabilities to events. In the classical method of calculating probabilities, we assume a finite sample space consisting of N points. The collections of all the points in the space is described as a set $X = \{x_1, x_2, ..., x_N\}$. An event is any subset of X, say $A \subseteq X$. For two such events $A \subseteq X$ and $B \subseteq X$, $A \cup B$ and $A \cap B$ are also events. Atomic events are given by singletons.

Instead of the sample space of events, a complex Hilbert space of infinite dimensions is used in QT. A Hilbert space is an abstract vector space, which includes a complex inner product between any two vectors in the space. For simplicity, we deal with a finite dimensional Hilbert space here. A N-dimensional Hilbert space comprises N orthonormal basis vectors $X = \{|X_i\rangle, i = 1, ..., N\}$. The choice of basis is arbitrary and there can be any number of basis for a Hilbert space. Here $|X\rangle$ is the way to denote a vector in the Dirac notation [Dirac 1982]. An event A is defined not by the subset of vectors $X_A \subseteq X$, but rather by a subspace spanned by this subset. If A is an event spanned by $X_A \subseteq X$ and B is an event spanned by $X_B \subseteq X$, the intersection of the two events, also called the "meet" and denoted as $A \wedge B$, is given by the span of vectors in the subset $X_A \cap X_B$. Similarly, the union of the events, called the "join" and denoted as $X_A \vee X_B$, is given by the span of vectors in $X_A \cup X_B$. Note how the set theoretical intersection and union of points are replaced by the span of the intersection and union of vectors. This structural property leads to the violation of the distributive axiom [Busemeyer and Bruza 2012]. Before talking about that further, we first discuss the concept of states and projectors in the quantum framework.

2.2.2 States of a Quantum System. In the classical framework, we have the concept of a probability distribution function $p(X_i)$, which assigns real numbers to each point X_i of a sample space. In the quantum framework, we define a state vector $|S\rangle$ of unit length in a Hilbert space X, which induces a probability distribution over the subspaces of the Hilbert space (Figure 3a). A subspace is represented in term of a projection operator P, which is Hermitian ($P\dagger = P$, where $P\dagger$ denotes the complex conjugate of the transpose of P) and Idempotent ($PP = P$). The probability induced by a state

vector $|S\rangle$ onto a subspace is given by the square of the projection of the vector onto the subspace. It is calculated as:

$$\begin{aligned} |P\,|S\rangle\,|^2 &= \langle S|\,P^\dagger P\,|S\rangle \\ &= \langle S|\,P\,|S\rangle \end{aligned} \tag{2}$$

Figure 3(b) shows a two-dimensional Hilbert space with state vector $|S\rangle$ projected onto a one-dimensional subspace, A_2. In this case the projector is given by $P_{A_2} = |A_2\rangle\,\langle A_2|$ and the probability distribution of the state given by the vector $|S\rangle$ is:

$$\begin{aligned} |P_{A_2}\,|S\rangle\,|^2 &= \langle S|\,P_{A_2}\,|S\rangle \\ &= \langle S|A_2\rangle\,\langle A_2|S\rangle \\ &= |\,\langle A_2|S\rangle\,|^2 \end{aligned} \tag{3}$$

Here the quantity $\langle A_2|S\rangle$ is the probability amplitude of the state $|S\rangle$ for the event A_2. The state of a quantum system $|S\rangle$ is in general a superposition of all possible events (see the next subsection for a discussion on superposition). As discussed before, the events are given by all the vectors of an orthonormal basis. In the basis $\{|A_1\rangle, |A_2\rangle\}$, the state of the system is represented as:

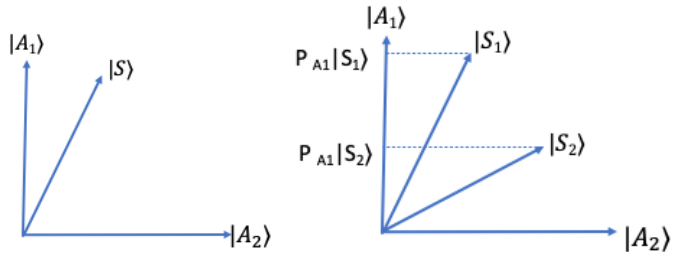
$$|S\rangle = a_1\,|A_1\rangle + a_2\,|A_2\rangle \tag{4}$$

where $a_1 = \langle A_1|S\rangle$ and $a_2 = \langle A_2|S\rangle$ are the probability amplitudes and $|a_1|^2$, $|a_2|^2$ represent the probabilities for events A_1 and A_2 to occur for the state $|S\rangle$. Hence $|a_1|^2 + |a_2|^2 = 1$.

2.2.3 Superposition and Collapse of a Quantum State. In models based on the classical probability theory, like Bayesian networks, a state of a system evolves from "moment to moment, but any given point of time the system is in a definite state" [Busemeyer and Bruza 2012]. To deal with uncertainty about which state the system is in, probabilities are assigned to each state. Thus, a dynamic system is in a definite state at each point of its evolution and is governed by a probability distribution over the states.

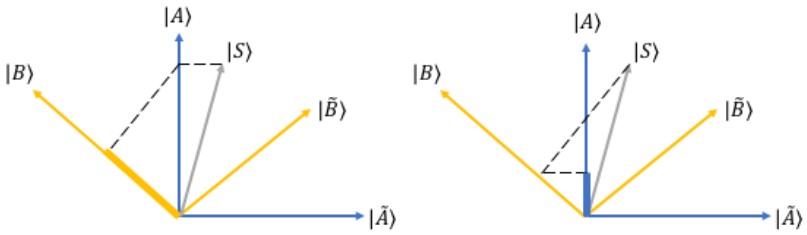
A quantum system is different from classical systems because of its ability to be in a superposition of all the possible states at the same time. This superposed state is a new state, which is not the same as any of the possible states of a classical system. Rather, it encapsulates the possibilities of being in all possible states. When a measurement is performed on a quantum system to learn its state, the superposed state collapses into one of the possible states with a certain probability. As an example, the system with state $|S\rangle$ in Figure 3(a) is a superposition of the basis vectors $|A_1\rangle$ and $|A_2\rangle$. It is neither in state $|A_1\rangle$ nor in $|A_2\rangle$. It is a new state with probabilities $|a_1|^2$ and $|a_2|^2$ for the system to be in state $|A_1\rangle$ or $|A_2\rangle$ upon measurement. Changing the probability amplitudes a_1 and a_2 leads to a new state, different from $|S\rangle$. This concept of collapse of a superposed state into one of the constituent states upon measurement is referred to as the Copenhagen Interpretation of Quantum Theory.

2.2.4 Violation of Distributive Axiom. In the classical probability theory, for a sample space $X = \{A, B\}$, the distributive axiom states that "$A = A \cap (B \cup \widetilde{B}) = (A \cap B) \cup (A \cap \widetilde{B})$, where $\widetilde{B}$ is the complement of event B. This axiom leads to the law of total probability" [Busemeyer and Bruza 2012]:



(a) (b)

Fig. 3. A two dimensional Hilbert Space with initial state vector and its projection



(a) (b)

Fig. 4. Two basis representing incompatible events showing order effects

$$\begin{aligned} p(A) &= p(A \cap X) = p(A \cap (B \cup \widetilde{B})) \\ &= p((A \cap B) \cup (A \cap \widetilde{B})) \\ &= p(A \cap B) + p(A \cap \widetilde{B}) \\ &= p(B)p(A|B) + p(\widetilde{B})p(A|\widetilde{B}) \end{aligned} \tag{5}$$

In simple terms, this law states that if an event A occurs, it can occur along with B or without B. In the quantum framework, consider a two dimensional Hilbert space with two basis vectors $|A_1\rangle$ and $|A_2\rangle$, as in Figure 3a. The intersection and union of two subspaces (called as *meet* and *join*, respectively) is defined as the intersection and union of the set of vectors spanning the subspaces, respectively (Section 2.2.1). Denoting the one-dimensional subspaces of this Hilbert space by their projectors P_S, P_{A_1}, P_{A_2}, the meet of the subspaces $P_S \wedge P_{A_1} = 0$ and $P_S \wedge P_{A_2} = 0$. The meet of two subspaces thus works the same way as intersection in set theory. The difference comes from the definitions of union and complement. The union or join of the two subspaces $P_{A_1} \vee P_{A_2}$ is the whole two-dimensional Hilbert Space, not just the set of two vectors $|A_1\rangle$ and $|A_2\rangle$ as in the set theory. Thus we get

$$P_S \wedge (P_{A_1} \vee P_{A_2}) = P_S \tag{6}$$

which violates the distributive axiom, as $(P_S \wedge P_{A_1}) \vee (P_S \wedge P_{A_2}) = 0$.

2.2.5 Compatible and Incompatible Events. Classical systems follow the principle of unicity [Griffiths 2001], which states that there is always a single sample space "which provides an exhaustive description of all the events that can happen in an experiment" [Busemeyer and Bruza 2012]. Therefore a single probability distribution function is sufficient to calculate the probabilities for all the events.

In the quantum framework, a state vector is represented as a superposition of all the basis vectors. One can choose to represent this state vector in any arbitrary basis. Thus the same state vector is expressed in different basis and each basis represents a particular property of the quantum system. The state vector induces different probabilities onto different basis of the Hilbert space. The state vector is thus an abstract entity. It does not have any fixed representation. A particular representation conceptualizes when we talk of a particular basis.

In Figure 4, we show a Hilbert space with two basis. One with orthonormal vectors $|A\rangle$ and $\left|\widetilde{A}\right\rangle$ and another basis with orthonormal vectors $|B\rangle$ and $\left|\widetilde{B}\right\rangle$. Consider the following events, in a particular order - A and B. To calculate the probability that these two events occur, the state vector $|S\rangle$ is projected onto the vector $|A\rangle$ and the new collapsed state is projected onto the vector $|B\rangle$. Hence we get the probability for A and B to occur as $p(A,B) = |P_B P_A |S\rangle|^2$, and using Equation 3, we get

$$p(A,B) = |\langle B|A\rangle|^2 . |\langle A|S\rangle|^2 \tag{7}$$

Now, if the same two events occur in the reverse order, i.e., B and then A, then the probability of them occurring is given by $p(B,A) = |P_A P_B |S\rangle|^2$, which, using Equation 3, is

$$p(B,A) = |\langle A|B\rangle|^2 . |\langle B|S\rangle|^2 \tag{8}$$

Now, Equations 7 and 8 will assign different values to the left-hand side when the value of the terms $\langle B|S\rangle$ and $\langle A|S\rangle$ are different. Which is the case if A and B are vectors in different basis. In the classical theory, we can assign joint probability distributions to two events occurring together regardless of their order, i.e., $p(A,B) = p(B,A)$, but such a joint probability distribution does not exist for events belonging to two different basis in a Hilbert space. We call these events as *incompatible events*. In the language of linear algebra, the projectors corresponding to these events do no commute, i.e., $P_A P_B \neq P_B P_A$. A geometrical explanation can be obtained from Figure 4. In Figure 4a, the order of projections is $S \rightarrow A \rightarrow B$ and in Figure 4b, it is $S \rightarrow B \rightarrow A$. We can see that the final projections (indicated by thick blue lines) in the two cases are different (and hence different probabilities) and depend upon the geometry of the Hilbert space (specifically, the angle between the vectors).

2.2.6 Density Matrices and Trace Rule. Another way of representing a quantum state, apart from the vector representation, is the density matrix or the density operator ρ. For a state $|\psi\rangle$, the density matrix is given by

$$\rho = |\psi\rangle\langle\psi| \tag{9}$$

which is a square matrix. The probability induced by the state represented by ρ onto a subspace represented by the projector P follows from the Gleason's Theorem [Gleason 1957], and is given by

$$Pr = tr(\rho P) \tag{10}$$

where $tr(x)$ is the trace of a matrix x, i.e. the sum of its principle diagonal elements. If we denote $P = |\phi\rangle\langle\phi|$, then the trace can be written as $tr(\rho P) = tr(\rho |\phi\rangle\langle\phi|) = \langle\phi|\rho|\phi\rangle$ which, for $\rho = |\psi\rangle\langle\psi|$ is

$$\mathrm{tr}(\rho P) = |\langle\psi|\phi\rangle|^2 \tag{11}$$

This is the same probability as calculated using the Born rule described earlier.

Density Matrix gives us the advantage of representing a mixture of classical and quantum systems. For example, if there is a mixture of n quantum systems, with the probability of each system being present in the mixture denoted by p_i, then this mixed system can be represented by a density matrix:

$$\rho = p_1\rho_1 + p_2\rho_2 + \ldots + p_n\rho_n \tag{12}$$

where ρ_i is the density matrix of the i-th quantum system, which has a classical probability p_i to belong to the mixture.

2.2.7 Composite Quantum Systems. Multiple quantum systems can be considered as a single system by combining their Hilbert spaces using a tensor product of the individual Hilbert spaces. If $|S\rangle_1, |S\rangle_2, ..., |S\rangle_n$ represent the states of n distinct quantum systems, the state of the composite quantum system of all these individual systems is given by $|S\rangle_1 \otimes |S\rangle_2 \otimes ... \otimes |S\rangle_n$, also denoted as $|S\rangle_1 |S\rangle_2 .. |S\rangle_n$. For example, consider two quantum systems represented by two dimensional Hilbert spaces with $|A\rangle$ and $|\tilde{A}\rangle$ as the basis vectors. Note that this is different from the system represented in Figure 3(b). Instead of multiple states in one Hilbert space, here we have multiple Hilbert spaces each with a state $|S\rangle_i$. The state of each of the systems (identical Hilbert spaces in this case) are given by: $|S\rangle_1 = \frac{1}{\sqrt{2}}|A\rangle + \frac{1}{\sqrt{2}}|\tilde{A}\rangle$ and $|S\rangle_2 = \frac{1}{\sqrt{2}}|A\rangle + \frac{1}{\sqrt{2}}|\tilde{A}\rangle$. Then the composite system is given by:

$$|S\rangle_1 \otimes |S\rangle_2 = \frac{1}{2}(|A\rangle|A\rangle + |A\rangle|\tilde{A}\rangle + |\tilde{A}\rangle|A\rangle + |\tilde{A}\rangle|\tilde{A}\rangle) \tag{13}$$

The above composite state is a separable state. It can be factorized into two separable components as: $(\frac{1}{\sqrt{2}}|A\rangle + \frac{1}{\sqrt{2}}|\tilde{A}\rangle) \otimes (\frac{1}{\sqrt{2}}|A\rangle + \frac{1}{\sqrt{2}}|\tilde{A}\rangle)$. There exists some composite systems for which it is not possible to separate the composite state back into single systems. A famous example of such states are Bell states:

$$|S^{\pm}\rangle = \frac{1}{\sqrt{2}}(|A\rangle|\tilde{A}\rangle \pm |\tilde{A}\rangle|A\rangle) \tag{14}$$

These states are called Entangled states and this property of Entanglement is a unique and a fundamental feature of Quantum Physics. When a measurement is performed on one part of the entangled system, the state of the other system can be known instantaneously, even if the two individual components are separated by a large distance. For example, consider two experimenters Alice and Bob who possess quantum states which are entangled with each other: $|S\rangle = \frac{1}{\sqrt{2}}(|A\rangle_1|\tilde{A}\rangle_2 + |\tilde{A}\rangle_1|A\rangle_2)$, where subscripts 1 and 2 denote that the states are possessed by Alice and Bob respectively. Now initially both the systems are in a superposition state. One cannot tell if it is in state $|A\rangle_i$ or state $|\tilde{A}\rangle_i$ ($i \in 1, 2$). If Alice measures her system and it collapses to, say, state $|A\rangle_1$ (it can collapse to either $|A\rangle_1$ or $|\tilde{A}\rangle_1$ with equal probability), then the state of the composite system collapses to state $|A\rangle_1|\tilde{A}\rangle_2$. Alice can instantaneously know that Bob's state has collapsed to state $|\tilde{A}\rangle_2$.

3 QUANTUM THEORY INSPIRED INFORMATION RETRIEVAL

The application of QT to Information Retrieval (IR) can be broadly divided into two major aspects. The first is the Quantum-inspired representation of entities like documents, queries, etc. in IR. Related to the representational aspect is that of Ranking in IR. The way the documents and queries are represented often determines the method for ranking documents. The second aspect is User interactions, including relevance feedback, query expansion, and user cognitive modeling. Projection models using the Hilbert space and multiple basis to represent states are used in both aspects: in representation - for abstraction of documents and queries, and in modeling user's dynamic and contextual information needs. Representation and user interaction areas in Quantum-inspired IR can be further sub-divided based on the specific approaches and applications (Figure 5).

3.1 Representation and Ranking

3.1.1 Subspace Representation and Projection Models. The first ideas regarding using the mathematical framework of QT in IR were described in the van Rijsbergen's pioneering book, *The Geometry of Information Retrieval* [van Rijsbergen 2004]. It addresses the need to develop a formal theory unifying different IR models, namely logic, vector space, and probabilistic models. It also sought to explore a formal description of user interactions and the abstraction of the concept

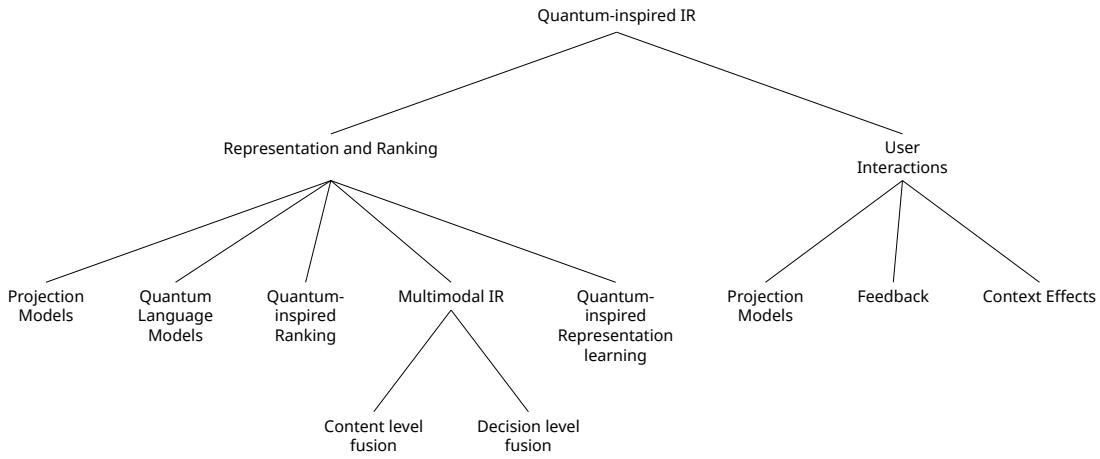


Fig. 5. Structure of the Quantum inspired IR survey

of a document in IR. Hence we find that the user is at center of most of the Quantum-inspired models and user interactions permeate all the representation and ranking methodologies, which we will discuss in the rest of this subsection.

Subspace Representation

A representation of document is usually related to the text it contains, but a document is in general a more abstract entity. To quote [van Rijsbergen 2004], "it is a set of ideas, a set of concepts, a story, etc." A document is defined as an abstract object that encapsulates answers to all possible queries. This is similar to a state vector in QT, which encodes information about all possible outcomes of measurement. The user interaction with an IR system is considered akin to measurement in QT, and the abstract document materializes to the specific information need of the user upon interaction. The Hilbert space representation of the Quantum framework is utilized to represent documents and queries in IR. It might seem similar to the Vector Space Model (VSM). However, instead of modeling documents as vectors in a term space, they are represented as subspaces of a concept space, spanned by a set of basis vectors.

Note that the documents and queries are themselves abstract and are defined in terms of the choice of basis. The same query or documents can be defined in different basis depending upon the user's point of view. The existence of multiple basis for the same state vector causes abstraction of objects in a Hilbert space. This, coupled with the fact that documents and queries are not merely vectors but subspaces in a complex, infinite dimensional vector space, gives us the leverage over the classical Vector Space Model. Besides providing a theoretical modification of the representational concepts of traditional IR, [van Rijsbergen 2004] also shows how the existing IR tasks like co-ordination level matching, feedback, clustering, etc. can be performed using the Quantum-like formulation.

Modeling queries and documents as multiple basis in IR was also investigated in [Melucci 2005a]. Documents and queries are modeled using certain semantic descriptors. However the semantic descriptors used for the same query or document may be different for different users, or different for same user in a different time, location or need. Therefore the use of descriptors depend upon the context. Since descriptors are modeled as basis vectors in a VSM, one can extend the VSM to include multiple basis, where each basis corresponds to a context. [Melucci 2005b] provides a method to

discover different contexts from data to model them as different basis, using a matrix decomposition algorithm (i.e., Cholesky's decomposition).

Information Need Spaces

A further development of the Quantum-inspired IR paradigm [Piwowarski and Lalmas 2009b] advocates the use of an information need space to model user interaction and evolving information need (IN) as part of representation. An IN is represented as a state in the form of a density matrix. For an ambiguous needs, the state is a mixed state, and if the IN is completely specified, it is a pure state. Before any user interaction, the IR system starts as a mixed state of all possible IN states. Consider the example when a user wants to order a pizza. In the beginning, the IN is in a mixture of all possible states, but a query "pizza" restricts the information need space to a subspace. Further interactions like knowing the time of the day, location of the user, etc. leads to smaller subspaces. Hence the evolution of information need is captured in the geometry. The representation of documents is proposed as in Structured Information Retrieval (SIR), which breaks away from representing the whole document as a single retrieval unit and uses document fragments like sections or paragraphs in response to a user query. It has been shown in [Piwowarski et al. 2008] that answers to queries may correspond to document fragments and not the whole document.

The specific details of building the information need spaces are given in [Piwowarski et al. 2010b]. The documents are modeled as a set of INs, with each IN being a vector. Using the SIR approach, documents are divided into fragments - paragraphs, sentences, sections or the document itself. Each document is converted into a vector using traditional techniques like tf-idf. Each of these fragments can satisfy an information need. Further, spectral decomposition of this set of vectors is performed to construct the document subspace. If the set of vectors for a document is U_d, then a subspace S_d comprises the span of the eigenvectors of the matrix $\sum_{u \in U_d} uu^T$. Only the eigenvectors corresponding to the top k eigenvalues are considered, since the low eigenvalues can be associated with noise.

[Piwowarski et al. 2010c] extend this work to include representation for queries. As a document is represented as a "set of pure IN vectors corresponding to different fragments of the document, a query term t is represented as a set U_t of IN vectors that correspond to document fragments containing the term t" [Piwowarski et al. 2010c]. Consider two documents D_1 and D_2 consisting of three different paragraphs each. Let $U_1 = \{v_1, v_2, v_3\}$ and $U_2 = \{v_4, v_5, v_6\}$ be the IN vectors corresponding to the documents. Taking the simpler case of a single term query, suppose the term occurs in paragraphs corresponding to the vectors v_2, v_5, v_6. Assuming that each fragment is equally likely to be a pure IN and a part of the user's actual IN, the density matrix for the query is written as:

$$\rho_q = \frac{1}{N_t} \sum_{\varphi \in U_t} \varphi\varphi^T \tag{15}$$

where $N_t = 3$ is the number of document fragments a term occurs in. Denoting the $S_d = \sum_{u \in U_d} uu^T$ as projector for a document as explained above, we can calculate the probability of relevance of the document for the query as:

$$p(Rel|q, d) = tr(\rho_q S_d) \tag{16}$$

For queries with multiple terms, either a weighted mixture of density matrices for the terms, or in an interesting case, the density matrix for a superposition of pure IN vectors can be used. Consider the three dimensional subspace of an information need space as shown in Figure 6. Let the vectors $\varphi_p, \varphi_{uk}, \varphi_{us}$ correspond to the INs "Pizza delivery", "Cambridge (US)" and "Cambridge (UK)" respectively. Then the IN for "Pizza delivery in Cambridge (UK)" would be represented by a superposition of φ_p and φ_{uk} vectors, as it is about both Pizza and Cambridge (UK). However the IN

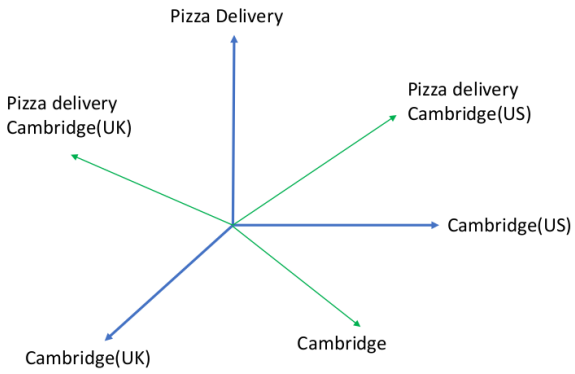


Fig. 6. Three dimensional Information Need (IN) space

"Cambridge" represents classical ambiguity regarding the country, and thus it is represented as a mixture of pure IN vectors φ_{uk} and φ_{us}. Thus a query "Pizza delivery in Cambridge" will be a mixture of superpositions.

This approach is later extended from a single Hilbert space of IN to multiple Hilbert spaces in [Piwowarski et al. 2010a]. User IN is considered to be composed of several "aspects", which need to be addressed by the relevant documents. Each aspect is represented in a separate Hilbert space made up of IN aspect vectors for the aspect. For example, consider a query "What tropical storms (hurricanes and typhoons) have caused significant property damage and loss of life?". It comprises two IN aspects: "tropical storms" and "significant damage/loss of life" [Piwowarski et al. 2010a]. The vectors for "hurricane" and "typhoons" are the IN aspect vectors for the tropical storm aspect of the query. As different aspect vectors belong to separate Hilbert spaces, the composite system corresponding to all IN aspects for the query is:

$$\varphi_q = \varphi_1 \otimes \varphi_2 \tag{17}$$

where φ_1 and φ_2 are constructed in the same way as Equation 15. The probability of relevance of a document defined by the subspace S_d in each Hilbert space would be $p(\otimes S_d|\varphi_q) = p(S_d|\varphi_1) \times p(S_d|\varphi_2)$.

The query representations for the above two approaches consider uniform weights to terms in mixtures and superpositions. However, the case of compound terms is not considered in [Piwowarski et al. 2010a]. The issue is dealt with in [Caputo et al. 2011], which provides a sophisticated representation of query density matrices. The paper introduces a query algebra, which can be "used to express relationship between query terms, thus allowing for more complex representations" [Caputo et al. 2011]. Several natural language processing (NLP) techniques such as Chunking and Dependency Parsing, are involved to identify different IN aspects and to characterize the relationship among terms within each aspect.

Polyrepresentation

The concept of representing information systems as composite systems in separate Hilbert spaces is explored in [Frommholz et al. 2010] for polyrepresentation of documents. A document may have different representations, based on different information sources, for example, text, author profiles, reviews and rating, etc. Each representation can correspond to different aspects of the information need of the user. Assume we have two Hilbert spaces representing a collection of books, one representing the authors and another for reviews. We have two authors $|Smith\rangle$ and $|Jones\rangle$

and two types of reviews $|Good\rangle$ and $|Bad\rangle$. Then a composite system of the two Hilbert spaces will be:

$$(|Smith\rangle + |Jones\rangle) \otimes (|Good\rangle + |Bad\rangle) = \tag{18}$$
$$|Smith\rangle\,|Good\rangle + |Smith\rangle\,|Bad\rangle + |Jones\rangle\,|Good\rangle + |Jones\rangle\,|Bad\rangle$$

where the user is uncertain whether to read a book by James or Smith and also unaware of their ratings. However, an interesting case is that of non-separable states, where a user wants a book by Smith which is rated good or wants a book by Jones which is rated as bad. The composite system of user's IN in this case is given by a non-separable state:

$$|Smith\rangle\,|Good\rangle + |Jones\rangle\,|Bad\rangle \tag{19}$$

which reduces the uncertainty from the system point of view.

3.1.2 Quantum inspired Language Models and Applications. The Quantum Language Model (QLM) proposed by [Sordoni et al. 2013b] combines the Vector Space Model and Probabilistic Language Model of classical IR via the Hilbert space formalism. The Quantum generalization of probabilities comes in the form of representing compound terms in queries and documents as superposition events, which have no classical analogue. This generalized Quantum probability model reduces to classical in case of using single terms only. More recently, a number of extensions of QLM have been made.

Basic QLM

In QLM proposed in [Sordoni et al. 2013b], a document or query is represented as a sequence of projectors. A projector represents a single term or compound term from the document/query. A document d containing words from a vocabulary of size N is represented as:

$$P_d = \{\pi_i : i = 1, ..., M\} \quad where \quad M \leq N \tag{20}$$

The Hilbert space is a term space of dimensionality N, where each vector $|v_s\rangle$ represents a term from the vocabulary. Thus the projector for a single term is $\pi_w = |v_s\rangle\langle v_s|$. The vector for a compound term $\left|v_{s_1..s_k}\right\rangle$ is the superposition of all the vectors corresponding to the single terms:

$$\left|v_{s_1..s_k}\right\rangle = \sum_{i=1}^{k} \sigma_i \left|v_{s_i}\right\rangle \tag{21}$$

where σ_i quantifies how much the compound term represents the single term s_i, and $\sum_{i=1}^{k} |\sigma_i|^2 = 1$. Thus in the same subspace, the representation of new term is created. This is not possible in traditional Vector Space Models because for every new term, single or compound, one has to add a new dimension to the vector space. Representing compound terms as superposition events solves that problem. Also, the compound term and the single terms in it are not disjoint and are related by the $\sigma_i s$.

Practically, in order to construct the projectors for a document, the terms co-occurring in the document within a fixed window of size L are considered as compound terms. A language model is essentially a density matrix ρ, and for a document it is represented by projectors $P_d = \{\pi_1, \pi_2, ..., \pi_M\}$. It is obtained by maximizing the following function:

$$L_{P_d}(\rho) = \prod_{i=1}^{M} tr(\rho\pi_i) \tag{22}$$

The language model is estimated using a generalization of an Expectation-Maximization based algorithm, called the $R\rho R$ algorithm [Lvovsky 2004].

The language model for a query ρ_q can be estimated in a similar way. The relevance of a document for a query can be calculated using a generalization of the KL divergence method called *quantum relative entropy* or *Von-Neumann(VN)*

divergence [Umegaki 1962]. Given two language models ρ_q and ρ_d, the scoring function is:

$$\Delta_{VN}(\rho_q||\rho_d) = -tr(\rho_q \log \rho_d) \tag{23}$$

where $tr(x)$ denotes the trace of the matrix x. Experimentally, the QLM has been shown to outperform a baseline language model and a Markov random fields (MRF) based model [Metzler and Croft 2005] (which was state-of-the-art at the time of publication of [Sordoni et al. 2013b]) for document ranking.

Extended QLMs

Several extensions have been made to the basic QLM described above. [Xie et al. 2015] propose to augment the QLM by making use of "entangled" terms. Based on the relation between Unconditional Pure Dependence (UPD) and Quantum entanglement [Hou and Song 2009; Hou et al. 2013], the UPD patterns are extracted from queries and documents, and the corresponding projectors are constructed. Instead of considering arbitrary compound terms, these UPD patterns are used as they show a statistical relationship similar to entangled systems.

Moreover, the $R\rho R$ algorithm used in QLM has a disadvantage in that it does not always converge. Hence a new global convergence algorithm is used in [Zhang et al. 2018b] for a Global Quantum Language Model (GQLM) but applied on twitter sentiment analysis tasks. Two dictionaries of positive and negative sentiment words are constructed. The global convergence algorithm constructs density matrices for the dictionaries and documents. Then using the quantum relative entropy, a document is projected onto each dictionary to determine its sentiment class.

A Quantum language model based Query expansion approach is presented in [Li et al. 2018a]. Using the GQLM described earlier, the language models for documents and query are constructed. The initial ranking is achieved using the Quantum relative entropy. Then a density matrix is constructed for the top k retrieved documents. The top n non-query terms, corresponding to the top n diagonal elements of the density matrix, are selected as expanded terms. The top n diagonal elements are in the order of the Quantum probabilities of the terms. Hence the advantage of using Quantum probabilities can be intuitively understood from here - the quantum probability reflects both the single term occurrence and the co-occurrence between terms. Hence, "a term with a high frequency but a low co-occurrence with other terms may as well have a lower quantum probability than a terms with lower frequency but higher co-occurrence" [Li et al. 2018a]. After having formed the expanded query, a new GQLM is constructed for the expanded query and the documents are re-ranked accordingly. Experiments on the TREC 2013 and 2014 session track datasets show a better performance than the original QLM and another quantum model proposed in [Wang et al. 2018], which is based on user interactions. Indeed, there have been various extensions of QLM that adapt to user interactions [Li et al. 2016, 2015; Wang et al. 2017], which will be discussed in Section 3.2.

The QLM is also extended within a neural network structure in [Zhang et al. 2018a] for Question Answering (QA), while the authors mention that the model can also be applied to other IR tasks, such as ad-hoc retrieval. Using word embeddings as vectors, a density matrix for each sentence is constructed, for both questions (as queries) and answers (as documents). The density matrix represents a mixture of semantic spaces. A joint representation of queries and documents is constructed by multiplying the density matrices for queries and documents. Then a convolution layer is applied over this joint representation followed by pooling, a fully connected layer and a softmax layer. The binary output of the softmax layer represents probabilities of relevance and non-relevance of the answer with respect to the question. This process is repeated for each question and answer pair and a ranking based on their relevance probabilities is produced. This model achieved MAP and MRR scores of 0.7589 and 0.8254 on the TREC-QA dataset, which was 2.46%

and 3.24% improvement over a neural model TANDA [Yang et al. 2016a], which was state-of-art at the time when the paper was published[3].

A Quantum many body wave function based language model is presented in [Zhang et al. 2018d]. The aim is to create a language model which addresses the challenges in word combinations, where each of the individual words can possess multiple meanings. Different meanings of a word are represented as different basis vectors of a Hilbert space. The state vector for a word is a superposition of different base vectors corresponding to different meanings of the word. The state vector of sentence is represented as a tensor product of the individual word's state vectors. This is termed as a local representation, and a similar global representation of the language model is constructed using another corpora, to account for unseen words and unseen compound words. The global representation is projected onto the local representation akin to the smoothing process in classical language models. As the global representation is a higher rank tensor, it is decomposed using tensor decomposition techniques. The projection of the reduced tensor onto the local representation tensor is realized in the form of a convolutional neural network. While it is unable to outperform the above mentioned model [Zhang et al. 2018a] on the TREC-QA dataset, it performs significantly better on the WikiQA dataset[4].

Other Quantum-inspired language models

In [Blacoe et al. 2013], the Quantum theoretic framework is used to construct a syntax-aware semantic model. It also takes word order into account, unlike the QLM. Firstly, for each sentence, dependency parsing is performed and a set of dependency relations are extracted. This set is partitioned into clusters of syntactically similar relations, and each cluster is assigned a Hilbert space. Each Hilbert space has the word vectors as the basis vectors. The state vector in each Hilbert space is a superposition of the word vectors, which are dependencies of the same word. The state vector for a given word is written as a tensor product of the state vectors in all the Hilbert spaces. A complex phrase is ascribed to the state vector. A word occurring in different senses will have different state vectors, which are then superposed to get the final vector for each word. It is then converted into a density matrix, and the density matrices of the occurrences of the word in different documents are added up. The similarity between two words can be measured using the trace rule, which essentially takes the pair-wise inner products of the state vectors. This allows the words to "select" each other's context and should lead to more accurate similarity values. Experiments done on word similarity and word association tasks reveal a better performance than some classical models. This method is extended in [Blacoe 2015] to construct density matrices for sentences. To create a density matrix for a sentence, first the dependency parsing tree is constructed. For each node in the tree, its dependencies are projected onto it and the post projection states are summed up together with the density matrix of the node. This procedure is performed recursively until the whole sentence is covered. The method is tested on the paraphrase detection task with the Microsoft Paraphrase Detection dataset, and shows better accuracy and F1 scores than a recent neural network model reported in [Shen et al. 2018].

In [Basile and Tamburini 2017], an n-gram language model inspired from QT is introduced, with application to speech recognition. Unlike the quantum inspired language models presented earlier, this paper makes use of the unitary evolution of a quantum state in time, e.g. $\rho_{t+1} = U\rho_t U^\dagger$, where U is a unitary operator which changes the state of a system ρ_t. To measure the probability of a word w, the system state is projected onto the state of the word w using the projector $\Pi_w = |w\rangle \langle w|$. Probabilistic information about a sequence of words $w_1, w_2, ..., w_n$ is encoded in a density

[3]Note that the current state-of-the-art BERT-based neural model for TREC-QA has achieved MAP and MRR of 0.943 and 0.974 respectively [Garg et al. 2019].

[4]The TANDA model mentioned above currently gives the best performance on WikiQA dataset (MAP and MRR of 0.92 and 0.933 respectively, as compared to 0.695 and 0.71 by [Zhang et al. 2018d]).

matrix built using the following process:

$$\begin{aligned} p(w_1; \rho_0, U) &= tr(\rho_0, \Pi_{w_1}) \\ \rho_1^{'} &= \frac{\Pi_{w_1} \rho_0 \Pi_{w_1}}{tr(\Pi_{w_1} \rho_0 \Pi_{w_1})} \\ \rho_1 &= U \rho_1^{'} U^{\dagger} \end{aligned} \tag{24}$$

where ρ_0 is the initial state of the system, and ρ_1 is state of the system after processing the first word in the sequence. The unitary matrix U is responsible for the time evolution of the system and $tr(x)$ stands for the trace operation. The final probability of the whole sequence becomes:

$$p(w_i | w_1, ..., w_{i-1}) = tr(\rho_{i-1} \Pi_{w_i}) \tag{25}$$

One possible issue arises here. Continuously projecting and collapsing the system state to individual words may remove any quantum effects from the system, i.e. the system reduces to a classical markov model like system. To address the issue, the system is coupled with an ancillary system to avoid the complete collapse. A D-dimensional Hilbert space represents the ancillary system and thus the composite system has the Hilbert space $H_2 = H_{ancilla} \otimes H$. The new projectors for the composite space are given by $\Pi_w^{(2)} = I_D \otimes \Pi_w$. The advantage of doing this is that the time evolution of the composite system can give rise to non-trivial correlations between them (analogous to non-separability and entanglement) so that even when the state of the word sequence is collapsed, some information is retained in the ancillary part (owing to their non-trivial correlations). The words are represented in low-dimensional vectors and for each dimension, a unitary matrix is assigned for the composite system. The parameters are learned by minimizing the perplexity of the corpus of sentences. The perplexity is given by:

$$\Gamma(\rho_0, U) = exp(-\frac{1}{C} \sum_{w \in S} \log p(w | \rho_0, U) \tag{26}$$

Experiments on the TIMIT dataset show that this n-gram quantum language model has a lower perplexity than the state-of-the-art deep neural network architectures like RNN and RNN-LSTM. Although the paper reports an application of the proposed language model in speech recognition, it would be interesting to use it to construct document and query language models.

3.1.3 Quantum-inspired Ranking. The research on quantum-inspired ranking has been done from two perspectives: quantum probability ranking principle and quantum-like measurement.

Quantum Probability Ranking Principle

According to the Probability Ranking Principle [Robertson 1977], an IR system should rank the documents for a user IN in a decreasing order of their probability of relevance. It makes the assumption that "the relevance of a document to an information need does not depend on other documents" [Zuccon et al. 2009]. However, in real world situations, judgment of documents by a user is influenced by its previously judged documents [Eisenberg and Barry 1988]. "The utility of a document may become void if the user has already obtained the same information" [Zuccon et al. 2009]. This 'interference' between documents can be due to information overlap between documents or a change in the IN, and is accounted for in a Quantum Probability Ranking Principle (QPRP) [Zuccon et al. 2009]. QPRP draws an analogy [Melucci 2010] with the Double Slit Experiment by assuming the two slits to be two documents A and B which the user judges for a query. The position x on the screen corresponds to the event that the user is satisfied by the documents A and B, and decides to stop the search. If A is first document presented to the user, we have $p_{AB}(x)$ as the

probability that the user stops the search at document B. In the Double slit experiment, if slit A is fixed and slit B is varied in dimensions, which is analogous to having different documents listed after document A, we get $p_{AB_i}(x)$ as "the probability of stopping the search process after seeing documents A and B_i" [Zuccon et al. 2009]. The problem then boils down to finding which configuration of slits (set of documents) AB_i exhibits maximum value of $p_{AB_i}(x)$.

For the classical case, if there is no interference, i.e. only one of the B_i slit is opened at a time, we have "$p_{AB_i}(x) = p_A(x) + p_{B_i}(x)$" [Zuccon et al. 2009]:

$$\arg\max_x(p_{AB_i}(x)) = \arg\max_x(p_A(x) + p_{B_i}(x)) = \arg\max_x(p_{B_i}(x)) \tag{27}$$

However, in the quantum case, with all slits open, or all documents considered by the user till B_i, $p_{AB_i}(x) = p_A(x) + p_{B_i}(x) + I_{AB_i}(x)$, where $I_{AB_i}(x)$ is the interference term. Thus:

$$\begin{aligned}\arg\max_x(p_{AB_i}(x)) &= \arg\max_x(p_A(x) + p_{B_i}(x) + I_{AB_i}(x)) \\ &= \arg\max_x(p_B(x) + I_{AB_i}(x))\end{aligned} \tag{28}$$

Hence the best choice of document to rank after A is not the one whose relevance probability is maximum, but rather the one whose sum of individual relevance probability and the interference term with A is maximum. Hence, between two documents B and C, B is ranked before C if and only if:

$$p_B(x) + I_{AB} \geq p_C(x) + I_{AC}(x) \tag{29}$$

Recall from Equation 1 that the interference term depends upon the phase difference of the probability amplitudes of two quantum systems. Thus:

$$p_{AB}(x) = p_A(x) + p_B(x) + 2\sqrt{p_A(x)}\sqrt{p_B(x)}\cos(\theta_{AB}) \tag{30}$$

In [Zuccon et al. 2009], the authors did not give details of how to estimate the interference term. This estimation is done in an application of the QPRP to subtopic retrieval [Zuccon and Azzopardi 2010]. Subtropic retrieval is a task of providing a list of documents which covers all possible topics (IN aspects) relevant to the user IN. It advocates a more diverse ranking of documents, to achieve a minimal redundancy. Thus redundant documents can be assumed to be destructively interfering (negative interference term) and the documents having exclusive information be positively interfering. The $\cos(\theta)$ part of the interference term is estimated as the Pearson's correlation between the term vectors of two documents. The term vectors are constructed using the BM25 scheme. Experiments show that the QPRP based ranking for subtopic retrieval performs better than a model based on Portfolio Theory[Wang and Zhu 2009] (then state-of-the-art) for subtopic retrieval on the ClueWeb09-B collection.

Quantum Measurement inspired Ranking

Another method for document ranking using Quantum probabilities is discussed in [Zhao et al. 2011], called Quantum Measurement inspired Ranking (QMR). Document retrieval process is considered to be similar to a photon polarization process. A photon has a Horizontal or Vertical polarization, which can be measured by a polarizer. There also exist superposition states of both vertical and horizontal polarizations, which is detected by a horizontal or vertical polarizer rotated at a 45 degree angle.

$$|\nwarrow\rangle = \frac{1}{\sqrt{2}}(|\uparrow\rangle + |\downarrow\rangle) \tag{31}$$

Superposition states can be generated by passing a horizontal or vertically polarized photon through the rotated polarizer. Mathematically, the vertical and horizontal polarizers form an orthonormal basis of a two dimensional Hilbert space. The rotated polarization state forms another orthonormal basis in the same Hilbert space. In the analogy, the first round of document retrieval for a query is analogous to the measurement along the vertical or horizontal basis.

Then, a second round retrieval is performed to re-rank the documents by comparing all retrieved documents with the top k documents. This is analogous to passing the photons coming from a horizontal or vertical polarizer through the rotated polarizer. Mathematically this is formulated as projecting a vector represented in one basis onto the subspace generated by another rotated basis.

In the first round of retrieval, let $|\uparrow\rangle$ and $|\downarrow\rangle$ denote relevance and non-relevance of document respectively for a query. Then a document d with relevance probability $|\alpha_d|^2$ is represented in the first round as:

$$|d\rangle = \alpha_d |\uparrow\rangle + \beta_d |\downarrow\rangle \tag{32}$$

Taking the simple case of $k = 1$, let the topmost document in the first round of retrieval be represented as:

$$|t\rangle = \alpha_t |\uparrow\rangle + \beta_t |\downarrow\rangle \tag{33}$$

Then, re-ranking is done by representing the document d in terms of t:

$$|d\rangle = \lambda |t\rangle + \mu |\widetilde{t}\rangle \tag{34}$$

where $\lambda = \alpha_d\alpha_t + \beta_d\beta_t$ (see appendix in [Uprety and Song 2018] for a proof). The probability of relevance of the document d when re-ranking is performed using the top-ranked document of first round is the square of the projection of d onto t, which is $|\lambda|^2$, multiplied by the probability of relevance of t, which is $|\alpha_t|^2$:

$$p(d|t) = |\lambda\alpha_t|^2 \tag{35}$$

When $d = t$, then $\lambda = 1$ and the probability becomes $|\alpha_t|^2$, the original probability of relevance of the top-ranked document. The QMR approach performs significantly better than the QPRP on four TREC collections - WSJ9872, AP8889, ROBUST04 and WT10G on MAP ranking metric.

Quantum-inspired ranking has also been used to solve the query drift problem, which is defined as the inferiority of results obtained on query expansion, than the original query. This is because the underlying intent of the query might change upon expansion. Several solutions have been proposed for the query drift problem using pseudo relevance feedback [Zighelnic and Kurland 2008], focusing on the combination of document scores in the ranked lists of documents based on the original query and the expanded query. For example:

- CombMNZ rewards documents that are ranked higher in both original retrieval list and second retrieval list by adding the relative score of a document in each of the two lists.
- Interpolation technique makes a weighted addition of relative scores in the two lists.

In [Zhang et al. 2011], a document d is represented in terms of relevance and non-relevance for a query q:

$$|d\rangle = a_d |q\rangle + b_d |\widetilde{q}\rangle \tag{36}$$

where $|q\rangle$ and $|\widetilde{q}\rangle$ represent the vectors for relevance and non-relevance of d with respect to q, respectively. In terms of the expanded query q^e, the document is represented as:

$$|d^e\rangle = a_d^e |q^e\rangle + b_d^e |\widetilde{q^e}\rangle \tag{37}$$

"To prevent query drift, the existing fusion models in [Zighelnic and Kurland 2008] directly combine the two probabilities $|a_d|^2$ and $|a_d^e|^2$" [Zhang et al. 2011]. The CombMNZ reduces to:

$$(\delta_q(d) + \delta_q^e(d)).(\delta_q(d)|a_d|^2 + \delta_q^e(d)|a_d^e|^2) \tag{38}$$

where $\delta_q(d) = 1$ if d is relevant to query q. Similarly, the interpolation method becomes:

$$\lambda\delta_q(d)|a_d|^2 + (1-\lambda)\delta_q^e(d)|a_d^e|^2 \quad 0 \leq \lambda \leq 1 \tag{39}$$

However, the two probabilities $|a_d|^2$ and $|a_d^e|^2$ are under different basis and we need to write one in terms of the other. The Quantum Fusion Model (QFM) proposed in [Zhang et al. 2011] does that and the final outcome combines the probabilities in the following way:

$$(\delta_q(d)|a_d|^2).(\delta_q^e(d)|a_d^e|^2) \tag{40}$$

Thus the Quantum based model is a multiplicative model, while the classical models are additive. Another slightly modified version is:

$$(\delta_q(d)|a_d|^2).(\delta_q^e(d)|a_d^e|^2)^{1/\eta} \tag{41}$$

where "a small η can make scores of different documents retrieved for q^e more separated from each other, leading to more distinctive scores" [Zhang et al. 2011]. The QFM achieves a better performance than the CombMNZ and interpolation methods in terms of Mean Average Precision (MAP) of retrieved documents.

3.1.4 Multimodal Information Retrieval. Despite the wide application of QT in text-based IR, limited attention has been paid to multimodal IR, which is of increasing significance in many applications. The work in this area can be divided into feature level fusion and decision level modality fusion.

Feature Level Fusion

Initially, Wang et al. [Wang et al. 2010a] exploited tensor product of Hilbert spaces to fuse textual and image features for circumventing the heuristic combination of uni-modal feature spaces. In particular, textual and visual features are combined in a similar way as non-separable states of a Quantum system. The authors claim that the proposed modality fusion approach is able to capture cross-modal dynamics, i.e., interactions across different modalities (e.g., text and image modalities). In each single modality feature space, documents are formulated as a superposition of terms. These terms are words from a vocabulary for the text representation and visual words for the image representation. For instance, in the textual feature Hilbert space denoted as H_T:

$$|d_T\rangle = \sum_i w_{t_i} |t_i\rangle, \tag{42}$$

where the squared amplitude $w_{t_i}^2$ equals the probability of a document to be about the term t_i with $\sum_i w_{t_i}^2 = 1$. Similarly, for the image modality the formulation is as follows:

$$|d_V\rangle = \sum_i w_{v_i} |v_i\rangle, \tag{43}$$

where v_i represents visual words in the image Hilbert space H_V. Each pure state is modelled through a density matrix. In mathematical language, each density matrix is defined as the outer product of a superposition state. For example, for the textual representation, the density matrix is:

$$\rho_{d_T} = \sum_i p_i \left|d_{T_i}\right\rangle \left\langle d_{T_i}\right|, \tag{44}$$

where p_i is the probability of the state being in the basis state $\left|d_{T_i}\right\rangle$. Then, the text and image modalities are fused by taking the tensor product of the text and image Hilbert spaces as follows:

$$\rho_{d_{TV}} = \rho_{d_T} \otimes \rho_{d_V} + \rho_{correlation}, \tag{45}$$

where $\rho_{d_T} and \rho_{d_V}$ are the textual and visual density matrices respectively, and $\rho_{correlation}$ is the density matrix capturing cross-modal interactions between the text and image features. The resultant product is still a valid density matrix. Finally, for measuring the similarity between a multimodal document and query, the trace rule is used as follows:

$$sim(d, q) = trace(\rho_{d_{TV}} \cdot \rho_q), \tag{46}$$

where $\rho_{d_{TV}}$ and ρ_q are multimodal document and query density matrix representations respectively.

For capturing cross-modal interactions across the two modalities, two statistical approaches were explored: (a) the maximum feature likelihood that associates text with the maximal likely image features; and (b) the mutual information matrix that measures the mutual dependence between the two modalities. Experiments show that even without considering the correlation between text and image features, the pure tensor product approach outperforms other methods such as the use of image features or text features individually, or the concatenation of text and image features. However, such a method suffers from exponentially increasing computational complexity, as the outer product over multiple modalities results in high dimensional tensor representations. The experiments also show that the two proposed statistical methods are trivial to capture cross-modal interactions. For example, simple visual features were used, such as colour histograms, which can hardly be associated with high-level semantics. A more robust statistical approach, such as the cross-modal factor analysis [Atrey et al. 2010], might be more effective. Another issue was that images with limited or no annotation were lowly ranked or not retrieved at all. This implies that tensor product cannot manipulate missing values, which becomes a common problem in a real-world scenario. An automatic annotation task might circumvent the above problem. Also, assuming orthogonality of dimensions disregards any semantic overlap. This was an issue for textual space as the dimensions representing words need to represent language attributes such as polysemy and synonymy. Nowadays, we can address such issue by exploiting neural network language technologies [Devlin et al. 2019; Pennington et al. 2014] for constructing text vector spaces with compact semantic information.

Kaliciak et al. [Kaliciak et al. 2011] followed up with the previous model, aiming to solve the problem of missing modalities, e.g., when images are not annotated. They proposed two approaches to alleviate this problem, which can be easily integrated with the tensor-based fusion method. The first approach projects an un-annotated image onto a subspace generated by subsets of annotated images. In particular, by exploiting the Born rule, the square projection on the basis states results in a probability distribution of terms for each un-annotated image. The second approach alternatively utilizes the trace rule to calculate the similarity between an annotated and un-annotated image. Images are formulated as density matrices, entailing a probability distribution of terms. The results showed that such approaches under-performed the standard clustering techniques. The result might be related to the assumption that "the correlation at the image-level (i.e., images referring to the same topic) are stronger than the correlations based on the proximity between image terms (i.e., instances of image words)" [Kaliciak et al. 2011].

Later on, Kaliciak et al. [Kaliciak et al. 2013] proposed a quantum-inspired framework for a cross-modal retrieval task. That is, given a text query, to retrieve the most relevant images. They first constructed a common Hilbert space by taking the tensor product of image and text density matrices. Both text queries and image documents are represented in the joint Hilbert space. In this joint space, they also utilized a mechanism of trans-media pseudo-relevance for re-ranking retrieved images. Then, a projection measurement measures the relevance between the text query and each image document represented on the same space.

Decision-level Fusion

Decision level fusion combines uni-modal classification results to reach a final decision. Despite being a common approach for fusing different modalities, only preliminary studies have investigated quantum-inspired decision level approaches for IR tasks. Gkoumas et al. [Gkoumas et al. 2018] investigated non-classical correlations between mono-modal decisions on a pair of text-image documents for a multi-modal retrieval task. In principle, non-classical correlations

or quantum correlations are stronger than classical correlations due to latent contextual influences. In that study, the authors investigated the existence of this kind of non-classical correlation through the Bell inequality (CHSH inequality) violation in a small-scale experiment on the ImageCLEF dataset. Although they did not find a violation of the CHSH inequality, the experiment design provides useful insights for future investigations into such non-classical contextual correlations.

Quantum-inspired modality fusion models have also been developed for multimodal sentiment analysis. Sentiment is one of the factors considered by users in judging certain types of documents (e.g. news articles, blogs). Sentiment label can be considered as a feature in predicting relevance [Fuhr et al. 2018].

Zhang et al. [Zhang et al. 2018c] proposed a quantum-inspired decision level modality fusion approach for image-text sentiment analysis. Even this task is far from IR tasks, the approach could be fruitful for IR tasks as well. In particular, both text and image information is associated with density matrices which use the same globally convergent algorithm mentioned earlier in the case of extended QLMs to estimate the density matrices. In this way, density matrices capture the cross-modal interactions. Additionally, the human cognitive interference phenomenon caused when a user is exposed to conflicting text and image information channels, is also considered as analogous to quantum interference. Though, the interference term is treated as a single parameter and adjusted experimentally. The results suggest that, when the Cosine of the interference term equals 0.3, the model achieves the best performance. Moreover, the accuracy is the highest when a user pays more attention to the text instead of image modality, assigning weights 0.7 and 0.3 for the text and visual representation respectively. When the weights are reversed, the model performs the lowest. This is an interesting outcome since it helps us understand under which conditions the quantum-like interference works at the decision level and enhances explainability over the modality fusion process. Overall, large-scale experiments show that the proposed approach outperforms a wide range of state-of-the-art baselines.

A combination of Long Short Term Memory (LSTM) and a quantum-inspired framework for conversational sentiment analysis was proposed in [Zhang et al. 2019a]. In particular, words are represented as pure states in a real-valued Hilbert space. Then, a sentence is formulated as a mixture density matrix of pure states, i.e., unit vectors, which further is processed by a CNN, resulting in a dense representation. Next, the output of CNN is fed into an LSTM cell to make a decision. Considering conversation sentiment analysis contains time-series and thus requiring fusing time-varying signals, the authors exploited a sequence of LSTM cells and the concept of quantum-inspired measurement, namely weak measurement, to model inter-speaker sentiment influences over a dialogue.

[Zhang et al. 2020] is a follow up of [Zhang et al. 2019a] by extending the framework with two modalities, namely, text and visual modalities. Specifically, each modality is represented in an individual real-valued Hilbert space. The exact pipeline with [Zhang et al. 2019a] is followed to predict unimodal sentiment judgments. Then, the concept of quantum interference has been exploited to fuse text and visual sentiment judgements. Comprehensive experiments on two benchmarking datasets for conversational human language analysis showed that the proposed quantum-inspired framework beats the state-of-the-art performance for the video emotion recognition task. It is to be noted that conversational sentiment analysis is an important feature in conversational IR tasks.

3.1.5 Quantum inspired Representation Learning. The quantum-inspired representation and ranking models depend on the construction and learning of Hilbert spaces. These are developed or applied in areas like lexical semantic spaces, topic modelling, word embeddings, and text classification.

Lexical Semantic Spaces

The first connections between QT and semantic spaces were established in [Aerts and Czachor 2004]. In [Bruza and Cole 2005], one such connection is presented using the Hyperspace Analogue to Language (HAL) model [Burgess et al. 1998; Lund and Burgess 1996]. For a vocabulary of N words, the HAL algorithm constructs an $N \times N$ matrix by sliding a window of length l over a text corpus, thus capturing word co-occurrences within the window. Each element of the matrix measures word co-occurrence and in one way, word similarity. Each window is considered as a semantic space and approximates the context or the sense associated with the word. The semantic space for a word is computed in terms of the sum of semantic spaces. If there are y windows around the word w and x of them deal with a particular context i, then the semantic space S_i occurs with probability $p_i = \frac{x}{y}$ and the semantic space for word w can be written in terms of context semantic spaces as:

$$S_w = \sum_{i=1}^{m} p_i S_i \tag{47}$$

This formula is the same as that of a mixed density matrix written as a mixture of density matrices of pure states. Thus the context of words can be considered as pure states. HAL is also used in [Hou and Song 2009; Hou et al. 2013] to model word correlations like Quantum correlations of non-separable states.

The explicit term occurrence based approaches are insufficient to capture hidden semantics. The advent of machine learning techniques have opened up a door to learning semantic spaces based on topic modelling and word embedding.

Interference Topic Model

The analogy to Quantum interference is used in [Sordoni et al. 2013a] for modeling interactions between topics. Topic modeling is used to discover hidden themes in text collections. A topic is defined as a probability distribution over a vocabulary and a document is a mixture of one or more such topics. Finally, every word in a document is supposed to come from one of the topics. The probability of a term w in a document model θ_d with k topics is given as [Sordoni et al. 2013a]:

$$p(w|\theta_d) = \sum_k p(w|z = k, \phi) p(z = k|\theta_d) = \sum_k \theta_{dk} \phi_{kw} \tag{48}$$

where $w \in \{1, ..., N\}$ denotes a word from the vocabulary. $z \in \{1, ..., K\}$ is the index for a topic. θ_d denotes a document where $\theta_d = (\theta_{d1}, ..., \theta_{dk})$, θ_{di} being the "topic proportions for the document" [Sordoni et al. 2013a]. ϕ is a $N \times K$ matrix representing the distribution of topics over terms.

Consider the case of two topics: 'war' and 'oil'. The term 'Iraq' is present in both topics. Now if a document contains both topics, still the probability of term 'Iraq' in the document is less than the maximum of its probability in either of the topics. Mathematically speaking [Sordoni et al. 2013a],

$$\begin{aligned} p(w = Iraq|\theta_d) &= p(Iraq|war) * p(war|\theta_d) + p(Iraq|oil) p * (oil|\theta_d) \\ p(Iraq|\theta_d) &\leq \max(p(Iraq|war), p(Iraq|oil)) \end{aligned} \tag{49}$$

However, the probability of the term 'Iraq' occurring in the document should be significantly more given it contains topic 'war' and 'oil'. Current topic models do not consider the interference or relation between two topics when generating a word. They assume the topics to be independent. To capture topic dependence via Quantum probabilities, [Sordoni et al. 2013a] assumes a Hilbert space where each dimension corresponds to a word from the vocabulary. Then, each topic is a vector in this Hilbert space z_k, which is a superposition of vectors corresponding to the terms. Thus we have:

$$|z_k\rangle = \sum_w z_{kw} |e_w\rangle = \sum_w \sqrt{\phi_{kw}} e^{i\varphi_{kw}} |e_w\rangle \tag{50}$$

where $\sqrt{\phi_{kw}}e^{i\varphi_{kw}}$ is the complex amplitude for the topic $|z_k\rangle$ in state $|e_w\rangle$ and $|\sqrt{\phi_{kw}}e^{i\varphi_{kw}}|^2 = p(w|z = k, \phi)$. A document can be represented as a superposition of topic states, with the coefficients being the proportion of topic in the document.

$$|d\rangle = \frac{1}{Z_d}(\sum_k \sqrt{\theta_{dk}}\,|z_k\rangle) \tag{51}$$

where Z_d is a normalization constant. The projection of a document vector onto a word vector is given as:

$$d_w = \langle e_w|d\rangle \propto \sum_k \sqrt{\theta_{dk}\phi_{kw}}e^{i\varphi_{kw}} \tag{52}$$

The probability of a term in the document is given by:

$$p(e_w^+ e_w) = |\,\langle e_w|d\rangle\,|^2 \propto |\sum_k \sqrt{\theta_{dk}\phi_{kw}}e^{i\varphi_{kw}}|^2 \tag{53}$$

$$= \sum_k \theta_{dk}\phi_{kw} + 2\sum_{i<j}\sqrt{\theta_{di}\theta_{dj}}\sqrt{\phi_{iw}\phi_{jw}}\cos(\varphi_{iw} - \varphi_{jw})$$

This equation represents the proposed interference-topic model. The first part of the expression on the right hand side corresponds to the classical topic model given in Equation 48, and the second is "the interference term which boosts or penalizes the probability for term w in the final document model depending on the phase differences $\varphi_{iw} - \varphi_{jw}$" [Sordoni et al. 2013a]. For a particular word, if a pair of topics are in the same phase then, $\varphi_{iw} - \varphi_{jw} = 0$ and $cos(\varphi_{iw} - \varphi_{jw}) = 1$. This increases the probability of seeing the word w in the document. For the phase difference of $\frac{\pi}{2}$, the interference term vanishes and the classical topic model is recovered. In their experiment, [Sordoni et al. 2013a] estimated the interference term using a similarity measure between the topic distributions, such as the Cosine similarity. The topic model helps in relevance ranking in IR by providing a better match for queries and documents, beyond the term level. This Quantum-,inspired topic model is applied to retrieval tasks like the TREC newswire corpora and performs better than the classical topic model.

Complex Numbers

QT in its most general formulation uses complex numbers in its representations and computations. Without the use of complex amplitudes, for example, the interference effects will be restricted to only positive and negative interference values (+1 and −1), while not utilizing the full range of possibilities in between. Therefore, it is imperative that Quantum models outside of Physics which are directly or indirectly making use of the superposition and interference phenomena use complex numbers in representation of state vectors, in order to maximally exploit the power of quantum probabilities. However, it is difficult to get an intuition as to how to represent concepts, objects, terms, decisions, etc. using complex numbers.

[van Rijsbergen 2004] proposed using inverse document frequency (idf) of a term as the imaginary part and the term frequency (tf) as the real part of a complex number. In [Zuccon et al. 2011], this proposal was investigated and found to be performing poorly than the baseline Vector Space models. In [Wittek et al. 2014], different types of word semantics are combined using a complex Hilbert Space. The main idea is to represent distributional semantics, like the word co-occurrence information as real part and represent ontological information about words as the imaginary part of a complex valued vector. The real vectors are constructed using the technique of Random Indexing, where a word vector is constructed using the vectors that represent the contexts of the word. A document is then represented as a sum of its word vectors. Besides, using the technique of concept indexing, a document is also represented as a Bag of Concepts vector. Each word is mapped to one or more concepts from a medical ontology. These two representations are

merged into a single complex vector. Thus, similarity between two documents can be calculated as the inner product of the two complex vectors which will reflect both the distributional and ontological similarity. This model is used in the IR task of TREC Medical Records Track and the retrieval effectiveness is found to be better than either the term-based only or concept-based only approaches in terms of the Mean Average Precision (MAP) and Precision@10 metrics.

Quantum-inspired Neural Representation Models

In [Li et al. 2018b], the challenge of emergent meaning and sentiment of a combination of words is addressed. They hypothesize that humans do not associate a single meaning or sentiment to a word. A word contributes to the meaning or the sentiment of a sentence depending upon the other words it is combined with. For example, the words 'Penguin' and 'Flies' (Verb) might be neutral in polarity individually, but the phrase 'Penguin Flies' is of negative polarity. Similar examples can be constructed for sentiment of sentences. This is compared to the Quantum interference phenomenon where two superposed quantum states interfere and the final outcome depends upon their relative phase. As such a word embedding model using complex numbers is introduced. Each word is represented by a complex vector. It comprises a real part that holds word co-occurrence information, and a complex phase that captures abstract combinatory information like the sentiment factor. A sentence is considered as a superposition of words and thus a sentence vector can be constructed as a density matrix. This density matrix is learnable from labelled data, along with a projection matrix, which is used to calculate the probabilities assigned by the sentence vector. For a sentiment classification task, the projection matrix is used to classify the sentiment of the sentence according to the trace rule of calculating probabilities. The proposed model outperformed some word-embedding based models.

Wang et al. [Wang et al. 2019] proposed an end-to-end quantum-inspired neural framework for text classification. In particular, words are represented as pure states in a complex-valued Hilbert space, while sentences as mixture of pure states (i.e., words). Hence a sentence is represented in a mixture density matrix. In this work, phases are not defined explicitly but learnt through a backpropagation algorithm. Having said that, the exploitation of complex values is pivotal to the formulation of quantum concepts. Then, a set of measures is applied to the mixture density matrix, resulting in a sequence of scalar values. Practically, the measurements are related to high semantic concepts. Finally, a softmax function normalises the output of measurements into a probability distribution before classifying the sentence. Comprehensive experiments on six different datasets demonstrated the effectiveness of the proposed method against some neural models.

This framework was extended in [Li et al. 2019] for a question-answering task. The words are formulated as pure states in a complex-valued Hilbert space. Though, in contrast to [Wang et al. 2019], each word is represented in a pure density matrix while a sliding window is applied to the sentence, generating a local mixture density matrix for each local window. Both question and answer are represented by a sequence of mixture density matrices. Then, the same set of common measurements is applied to those density matrices in the sense that they share common semantic concepts. A max-pooling function is performed over the measurement output components, resulting in a dense representation before the matching process through the Cosine similarity measurement. The proposed complex-valued network for matching achieved comparable performances to strong CNN and RNN baselines on two benchmarking question answering (QA) datasets.

In [Zhang et al. 2018d], a quantum many-body wave function is used to model the semantic meaning of words within a local context, i.e., sentences, and a global context, i.e., corpus. In particular, each word is represented by different base states in the sense that each basis corresponds to a different word meaning. First, they model the word meaning within a sentence and corpus by the tensor product of basis states and then project the global tensor representation onto the

local one. This results in a high dimensional tensor, capturing interactions of words within a sentence and corpus. The high dimension resulted tensor representation is further decomposed in subspace base states, which finally processed by a CNN component for constructing the final representation.

Remark: The above works outperformed various existing neural models at the time when the papers were published. However, more recent neural models, such as BERT based models [Garg et al. 2019; Raffel et al. 2019] have achieved a largely improved performance on the same tasks. It is worth exploring the integration of quantum models with the new BERT architecture [Devlin et al. 2019] in the future.

Readers interested in tensor networks for representation learning, with or without quantum-inspirations, can also refer to [Zhang et al. 2019b] which introduces a Tensor Space Language model (TSLM) by building higher dimensional tensors using word vectors, leading to a generalisation of n-gram models.

In [Jiang et al. 2020], authors integrate quantum interference phenomena in neural networks with application to ad-hoc retrieval. Existing neural IR models are formulated in terms of classical probability. For example, if q_i represent sub-units (e.q. words) of a query Q, then neural models first perform sub-unit level matching and then aggregate the scores obtained to calculate a final relevance score. Formulated in terms of probabilities, such aggregation follows the law of total probability:

$$P(R_D|Q) = P(q_1)P(R_D|q_1) + P(q_2)P(R_D|q_2) \tag{54}$$

The authors hypothesise that a quantum-like cognitive interference can occur such that the aggregation of relevance scores can happen non-linearly, due to negative contribution of certain query or document sub-units. Thus the above equation becomes similar to Equation 1:

$$P(R_D|Q) = P(q_1)P(R_D|q_1) + P(q_2)P(R_D|q_2) + Int(R_D, Q, q_1, q_2) \tag{55}$$

They proceed to model this interference term and incorporate into a neural architecture. Query and document states are represented as a superposition of their respective sub-unit states with the coefficients of document sub-unit states being the tf-idf values, and those of query sub-units being trainable parameters. A composite system is constructed by taking the tensor product of the query and document state vectors. Then, calculating relevance probability using the trace rule breaks down the probability into two parts - similarity matching as used in neural IR models and the interference term, which is determined by the interaction between different document features. The similarity matching is achieved using a query attention mechanism, and the interference is modelled as a n-gram window convolution network. The new model is tested on Robust04 and Clueweb09-cat-B collections. While it performs better than various existing neural ranking models and even a vanilla-BERT [MacAvaney et al. 2019] (on P@20 metric on Robust04 dataset), it under-performs the current-state-of-the-art model [MacAvaney et al. 2019] on both NDCG@20 and P@20 on the Robust04. On the other hand, on the Clueweb-09-cat-B dataset it beats existing neural IR models and also the state-of-the-art XLNet model [Yang et al. 2019] in NDCG@20 and ERR@20 metrics.

Quantum-inspired Classification

Classification algorithms inspired by Quantum detection theory [Helstrom 1969] are discussed in [Buccio et al. 2018; Tiwari and Melucci 2018]. Binary classification is formulated as a signal detection problem. Projectors are defined to detect a signal, i.e. whether a document belongs to a topic or not. The average cost of incorrect detection (false alarm - detecting a signal when it is absent, miss - failing to detect a signal) is expressed in terms of projection operators and is minimized over a training set. Experiments performed over Reuters Newswire dataset show comparable results with Naive Bayes and SVM algorithms. In [Jaiswal et al. 2018], which used dimensionality reduction techniques for

word embeddings, the computation times for the complex embeddings in [Li et al. 2018b] is reduced, with additional application to the TREC-10 Question Classification task.

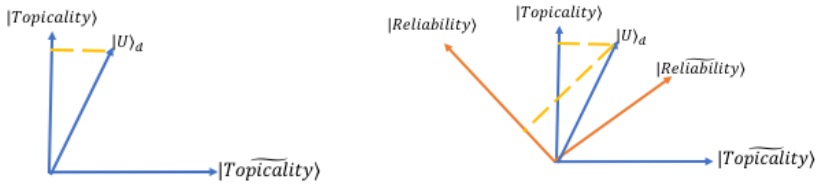
3.2 User Interactions

3.2.1 Projection Models. The area of user interactions in IR has many sub-aspects, primarily - the cognitive level of interaction, understanding the user IN by reformulation, expansion of queries, and building a user profile based on historical interactions. Earlier, we mentioned about the work in [Melucci 2005a,b], which use multiple basis of a Hilbert space to model different user contexts. This work is further extended in [Melucci 2008; Melucci and White 2007] to combine different user interaction and contextual features for Implicit Relevance Feedback (IRF). Their model uses interaction features like document display time, document saving, document bookmarking, webpage scrolling, webpage depth and document access frequency to construct a user interest profile. A basis vector represents each of these interaction features. The matching of documents against a user profile is done by projecting a document vector onto the subspace spanned by the basis vectors for the user profile. A large projection signifies high relevance of the document to the profile. The features described above are calculated for each document that the user has interacted with and a document-features correlation matrix is formed. Singular Value Decomposition (SVD) is performed to get the eigenvectors, which form the basis for a user profile.

In [Frommholz et al. 2011; Piwowarski and Lalmas 2009a], a general framework for query reformulation using Quantum probabilities is described. The queries are represented as density matrices in a term space and query reformulation updates the query density matrix, which can be used to detect change in user IN in a search session.

A Hilbert Space for user's perception of document relevance is constructed in [Uprety et al. 2018]. It deals with the challenge of modeling multidimensional relevance of documents. In an extended Multidimensional User Relevance Model (MURM) [Li et al. 2017; Zhang et al. 2014], seven factors or dimensions of relevance are identified, which influence user's judgment of a document. They are Topicality, Interest, Novelty, Understandability, Scope, Habit and Reliability. The features defined for each of these seven dimensions in [Li et al. 2017] are extracted from the retrieved documents of a query and fed into the LambdaMART [Burges 2010] Learning to Rank algorithm. Thus for each relevance dimension, a document has a relevance score. In other words, for a query one gets seven different ranking lists, one for each dimension. The scores obtained are converted into probabilities using max-min normalization. Representing user's relevance perception of a document with respect to each relevance dimension as a vector, and perception of non-relevance as its orthogonal vector, a document can be represented in a two dimensional vector space [Uprety et al. 2018]. The two relevance and non-relevance vectors for a relevance dimension form an orthonormal basis of the vector space. Figure 7a denotes one such vector for a user's perception of document relevance $|U\rangle_d$ with respect to Topicality dimension of relevance. The projection of the document perception vector on the Topicality vector is proportional to the probability of relevance of the document with respect to the Topicality dimension. Relevance and non-relevance of the document with respect to another dimension is represented as another set of orthogonal vectors, which, in general form another basis in the same Hilbert Space. In Figure 7b, we see that the projection of the document perception vector is different in the Reliability basis, suggesting that, while the document has a high relevance when considering the Topicality dimension, it is not so much relevant when considering the Reliability dimension.

The most popular interpretation of QT is that the state vector collapses upon measurement from a superposition to a definite state. When drawing an analogy, the user's cognitive state is generally considered as in a superposition of various Information Needs (IN) and on judging a document as relevant, it collapses to one particular IN. However, in practice, this may not always be the case. Even after judging documents, a user may still be in an ambiguous or



(a) Topicality dimension in a single basis (b) Topicality and Reliability in multiple basis

Fig. 7. Modelling user's perception of relevance dimensions in Hilbert space

superposed state and there may not have been an apparent change in the IN. The standard interpretation of state collapse may not accurately capture the evolution of IN. This challenge is investigated in [Wang et al. 2018] using a Quantum Weak Measurement (QWM) model. It is a generalization of the standard collapse model where the variance of measurement results is much larger. To test the weak measurement phenomena in user judgments, a study is carried out. Users are asked to judge documents on a -4 to 4 scale of relevance. For some query-document pairs, the users are asked to judge the same document for a second time. According to the standard collapse of the IN, after the judging a document as relevant, subsequent judgments will produce the same relevance result. However, it is found that in many cases, users change the relevance decisions. This will happen only when after the judging the document for the first time, the users were still uncertain about the document and their IN. The user's cognitive state was still superposed. This is especially the case where judgments on some documents are not trivial and difficult to make. The weak measurement model involves the Two State Vector Formalism (TSVF) of QT. In TSVF, the state of a system is not represented by a single vector $|\psi\rangle$ but rather by two vectors $|\phi\rangle$ and $|\psi\rangle$, where one vector represents the state of the system in the past (relative to a time t), and the other represents the state after time t. A weak measurement of an observable W on the system is given by:

$$w = \frac{\langle\phi| W |\psi\rangle}{\langle\phi|\psi\rangle} \tag{56}$$

This type of Quantum measurement is applied in case of session search. A user's IN is represented by two vectors. One contains the historical session information in terms of Implicit relevance feedback, and the other represents the current query, in terms of Pseudo Relevance Feedback. The document vectors, $|d_i\rangle$, are calculated using word embedding techniques, and then the corresponding projectors $P_{D_i} = |d_i\rangle \langle d_i|$ are constructed. The relevance probability of document d_i using weak measurement is calculated as:

$$p = \frac{\langle\phi_{past}|d_i\rangle \langle d_i|\phi_{curr}\rangle}{\langle\phi_{past}|\phi_{curr}\rangle} \tag{57}$$

The experiment is conducted on the TREC Clueweb12 document collection and the QWM method gives a better performance than the Quantum Language Model and its variations, as well as some state-of-the-art classical IR models.

The TSVF is also used in [Wang et al. 2017] to modify the query density matrix in QLM. A quantum state is denoted as $\langle\psi_{post}||\psi_{pre}\rangle$. Here $|\psi_{pre}\rangle$ is a state evolving from the past to the present and $\langle\psi_{post}|$ is a state devolving from future to the present. The previous user query in the session is considered for the past state and the current query for which the documents are to be retrieved is considered as the future query. Then, separate projectors are constructed for the past and future queries and the density matrix ρ_d for the document is estimated in the following manner:

$$\rho_d = \arg\max\left(\sum_{i=1}^{M_{past}} \log tr(\rho_d \Pi_i) + \sum_{j=1}^{M_{future}} \log tr(\rho_d \Pi_j)\right) \tag{58}$$

where M_{past} and M_{future} are the number of projectors (made up of single terms or compound terms) in the past query and the future query respectively.

3.2.2 Feedback. The query drift problem presented in the previous subsection is approached using user's search history in [Zhang et al. 2016]. A document is represented as a superposition of query vectors for current query and for a latent query defined by the user's query history.

$$|d\rangle = a_d\,|q_c\rangle + b_d\,|q_h\rangle \tag{59}$$

q_h denotes the user IN that the user implicitly has in mind based on historical context, but has not been explicitly expressed into words. A document, in the superposition state of being relevant to both the current (q_c) and latent query (q_h), is then evaluated in terms of an expanded query. This is similar to the double slit experiment analogy with the two slits representing q_c and q_h and the detector screen representing the evaluation of this document in terms of the expanded query. Thus the document relevance with respect to the queries q_c and q_h interfere with each other. If $|q_e\rangle$ represents the vector for the expanded query and $|d\rangle = a_d\,|q_c\rangle + b_d\,|q_h\rangle$, then the projection of document onto the expanded query vector is:

$$\begin{aligned} d \rightarrow q_e &= |\,\langle q_e|d\rangle\,|^2 \\ &= |a_d\,\langle q_e|q_c\rangle + b_d\,\langle q_e|q_h\rangle\,|^2 \\ &= |\,\langle q_c|d\rangle\,\langle q_e|q_c\rangle + \langle q_h|d\rangle\,\langle q_e|q_h\rangle\,|^2 \\ &= |\,\langle q_c|d\rangle\,\langle q_e|q_c\rangle\,|^2 + |\,\langle q_h|d\rangle\,\langle q_e|q_h\rangle\,|^2 + 2\,\langle q_c|d\rangle\,\langle q_e|q_c\rangle\,\langle q_h|d\rangle\,\langle q_e|q_h\rangle\cos\theta \end{aligned} \tag{60}$$

where θ is the phase between the two paths $d \rightarrow q_c \rightarrow q_e$ and $d \rightarrow q_h \rightarrow q_e$. We get interference between the two paths, because the actual path is superposed, $d \rightarrow (q_c \& q_h) \rightarrow q_e$. The first round retrieval is assumed to be using both the current and the latent query at the same time. This method of query expansion using user's previous interactions, is termed as the Quantum Query Expansion (QQE) approach for session search. It gives better results than the QFM discussed in the previous subsection, over the NDCG evaluation measure. [Li et al. 2015] propose a Session-QLM (SQLM) to model the evolving nature of user's IN in a search session. The evolution is modelled using density matrix transformation. The density matrix is constructed using user interaction features like clicked and skipped documents, dwell time, click sequence, etc. User's historical queries in context is also used in [Li et al. 2016] for a Contextual Quantum Language Model (CQLM). The density matrices are constructed to represent the language models for the current query and for the historical queries in a search session. They are then combined to give the CQLM.

$$\rho_{CQLM} = \xi \times \rho_c + (1 - \xi) \times \rho_h \tag{61}$$

where $\xi \in [0, 1]$ combines the two language models.

The construction of ρ_h is done by combining all the ρ_{h_i} of the previous queries in the session. The historical queries in the session which are similar to the current query are given more weight. Hence:

$$\rho_h = \sum_{i=1}^{n-1} \gamma_i \times \rho_{h_i} \tag{62}$$

where γ_i is the similarity between current query q_c and previous query q_i. The similarity is calculated by "representing each query as a TF-IDF vector, derived from the concatenation of all of its result documents." [Li et al. 2016].

The CQLM, however, was not designed to capture the evolution of user IN. To address this issue, the same paper further proposed an Adaptive CQLM (ACQLM) to model the evolution of user IN. The basic idea is to decompose the current query into three - a common part, an added part and a removed part, relative to the previous queries in

the session. For example, if $q_k = wxy$, $q_{k-1} = xyz$, "then xy is the common part, z is the added part and w is the removed part. The common part indicates the user's underlying search topic/theme for the session. The removed and added parts reflect the change in IN" [Li et al. 2016]. The ACQLM adjusts the QLM in such a way, as to assign a higher probability to the terms (or composite terms) of the common and added parts. Thus the ACQLM builds upon the CQLM by incorporating query change signals in a structured and intuitive way, moving the QLM into the right direction.

3.2.3 Context Effects. A series of research has been carried out from the user cognitive aspect of IR, drawing parallels from QT and using the Quantum framework to model and explain some of the aspects. An early work [Wang et al. 2010b] investigated the interference in relevance judgment of a topic caused by another topic. Consider the topics "Brave Heart" (William Wallace's nickname and the name for his film biography) and "William Wallace", and a biographical article about William Wallace. Both topics are relevant to the article. Consider another topic about "William Wallace's wife". In a user study, it was found out that when the topics "Brave Heart" and "William Wallace" were displayed together for the article, 93% users chose to judge the article as relevant to "William Wallace" and only 14% chose it as being relevant to the topic "Brave Heart". However, when "Brave Heart" was displayed together with "William Wallace's wife", 89% of the users judged "Brave Heart" as relevant to the article and 5% judged "William Wallace's wife" to be relevant. There were experiments conducted with different topics and articles and such type of context effects were found, where the presence of one topic or document influences the relevance judgment of another topic or document. In the first case, "William Wallace" is highly relevant to the article. It sets a high comparison baseline which affects the judgment for the topic "Brave Heart" and results in a low probability of relevance. However, it appears more relevant in comparison with "William Wallace's wife". For a Quantum probabilistic explanation of this result, we regard "William Wallace" and "William Wallace's Wife" as two different contexts for the topic "Brave Heart". Each context is described by a basis. So a document or topic d can be represented in the context basis as:

$$|d\rangle = a_1 |q_1\rangle + a_2 |\bar{q_1}\rangle \tag{63}$$

where $|\bar{q_1}\rangle$ represents the absence of context q_1. Representing a query q in the same basis as $|q\rangle = b_1 |q_1\rangle + b_2 |\bar{q_1}\rangle$, we can calculate the relevance of the document d for query q as:

$$\begin{aligned} p(d|q) &= |\langle q|d\rangle|^2 \\ &= (a_1 b_1 + a_2 b_2) * (a_1 b_1 + a_2 b_2)^\dagger \\ &= a_1^2 b_1^2 + a_2^2 b_2^2 + 2 a_1 b_1 a_2 b_2 \cos\theta \end{aligned} \tag{64}$$

where the probability amplitudes are complex quantities, and θ represents the phase difference term. The third term is the interference term, which can be positive or negative depending upon the phase differences. For some contexts, the interference term is negative and the relevance of the same document for the query can be low, which explains why "Brave Heart" is judged less relevant when seen in the context of "William Wallace". There is a negative interference term that lowers the probability of relevance for the given query/article.

A follow-on work [Wang et al. 2016] further explored the influence of context in document relevance judgment. It specifically investigates the presence of Order Effects in relevance judgment of documents. In the experiment, users are shown a pair of documents for a query and the relevance judgment by the user for a document is affected by the order, in which the document is presented. For example, for the query "Albert Einstein" users are shown documents about "Issac Newton" and "Theory of Relativity". The relevance probability of "Issac Newton" is lower when it is shown after "Theory of Relativity" (called a comparative context) than when it is shown first (non-comparative context). In simple terms, having seen a more relevant document first, user's perception of relevance for a particular document may

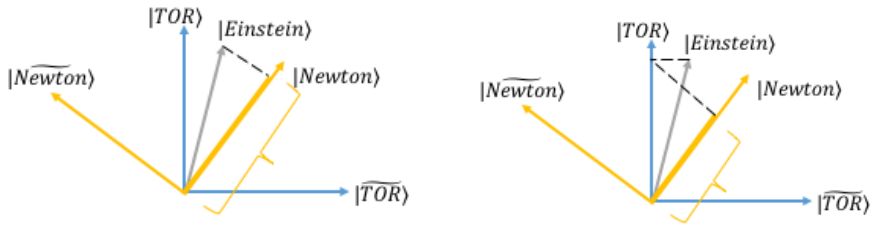


Fig. 8. On viewing document about Theory of Relativity, the judgment of topic Newton is lower for the query Einstein

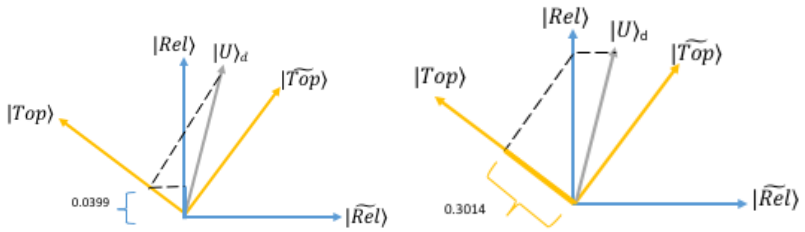
be lower. This can be explained as an Order Effect due to incompatibility between the topics, as shown in Figure 8. The paper also tested the Quantum Question Order inequality [Wang and Busemeyer 2013], which is an inequality for testing incompatibility in decision making systems.

One of the earliest works to investigate order effects in the different relevance dimensions is [Bruza and Chang 2014]. A user study was conducted, in which participants were asked questions about different pairs of relevance dimensions for a document, e.g. Credibility and Understandability, etc. It was found that the judgement of credibility, novelty, etc. was different depending upon the order in which they were asked to judge.

Similar order effects using query log data have been investigated in [Uprety and Song 2018]. The method of constructing a Hilbert space for multidimensional document relevance perception from [Uprety et al. 2018] is used (discussed earlier in this subsection). It is assumed that a user may consider multiple relevance dimensions while judging a document, for example, topicality and novelty. The relevance perception vectors corresponding to different relevance dimensions are in general incompatible in the Hilbert space representation. Thus different orders of consideration of the relevance dimensions may lead to different final judgment of the document. To investigate such behaviour in query log data, a subset of queries are found, where the top two retrieved documents have similar scores of relevance in all the seven dimensions. Yet the first document in the ranked list is not judged relevant, but the second one is. A small set of such queries are indeed found and order effects arising out of different order of consideration of relevance dimensions is offered as a possible explanation for such queries. Figure 9 explains the order effect for two documents d_1 and d_2 (ranking order for a query), which have the exact same Hilbert space, yet only d_2 is clicked. For d_1, if the user first considers Topicality and then Reliability to judge document d_1, then the final probability of judgment obtained is 0.0399 (Figure 9a). Whereas, for d_2, if the order is reversed, the probability of final judgment obtained will be 0.3014 (Figure 9b), much larger in this case. However, an important question is why the order is reversed in the user's mind for the next document. The authors argue that it could be due to a memory bias - the user can use the last relevance dimension considered for the previous document as the first dimension while judging the current document. Also, there is a possibility that such behaviour can be due to a variety of different reasons, or just random errors in the log data. Nonetheless, a quantum cognitive explanation based on order effects is a possibility.

As we see that there is some evidence of incompatibility between different relevance dimensions, [Uprety et al. 2019b] investigated the violation of Bell-type inequalities for multidimensional relevance judgment data. A violation of Bell-type inequalities would confirm the quantum nature of data. However, no such violation is found due to lacking probabilities of relevance available for joint judgment of a pair of documents in their dataset.

Order effect in risk and ambiguity is also investigated and observed in Information Foraging Theory in [Wittek et al. 2016]. This paper identifies a theoretical limit to simultaneous consideration of risk and ambiguity in decision making using eye tracking data, analogous to the uncertainty principle.



(a) User considers the sequence Topicality, Reliability while judging first document

(b) User considers the sequence Reliability, Topicality while judging second document

Fig. 9. Different order of consideration of dimensions leads to different final probability

In [Uprety et al. 2019a], the phenomena of incompatibility and order effects between relevance dimensions has been studied through a novel protocol design inspired from the Stern-Gerlach experiment of Physics. For a query-document pair, two groups of users were asked three questions relating to Topicality (T), Understandability (U) and Reliability (R) of a document, in orders TUR and TRU respectively. A complex-valued Hilbert space for the user cognitive state is constructed using the data obtained from the experiment, which is used to construct operators corresponding to the T, U and R measurements/judgements. Interference and incompatibility is discussed using these operators. This is the first work where complex numbers are used to capture interactions between relevance dimensions such as incompatibility and interference. It is extended in [Uprety et al. 2020] to test the violation of a Kolmogorovian probability axiom:

$$0 = \delta = P(A \vee B) - P(A) - P(B) + P(A \wedge B) \tag{65}$$

where the events A and B are the questions regarding Understandability and Reliability of a document. The conjunction and disjunction questions are asked through a specific experiment design and a violation of the above equality is observed in the data. Quantum model predicts a violation for all queries. This paper also compares quantum and Bayesian models for predicting multidimensional relevance probabilities. Quantum predictions are consistently closer to the experimental data, while predictions from the Bayesian model deviate significantly in some cases.

4 SUMMARY AND LIMITATIONS

van Rijsbergen's seminal work introduces us to similarities between the mathematical representation of microscopic particles in QT and information objects in IR. It does not delve into much depth over the distinct advantages of the quantum framework over traditional IR frameworks.

Early research inspired by van Rijsbergen's ideas implement QT-based ad-hoc IR models by considering information need space as Hilbert space and introducing ideas of superposition for ambiguous queries. These representations provide a good starting point in QR, but they generally fail to outperform the state-of-art methods in IR.

The Quantum Language Model (QLM) is a promising application and intends to solve a crucial problem in NLP and IR - of representing compound terms in relation to the individual terms. Superposition principle is made use of and a quantum algorithm to build a language model is applied. It performs better than baseline models like tf-idf and BM25. The later quantum-inspired language models show marked improvement over the QLM but need to be applied on a wide variety of IR and NLP tasks and compared with the state-of-the-art baselines. The complex word embedding is another promising approach, however there is a lack of clarity as to why this methods performs better than some classical methods and what is the intuition behind the interference terms and complex phases.

The Quantum Probability Ranking Principle is an important milestone in quantum-inspired IR as it approaches and combines QT and IR from an axiomatic point of view. However, the problem of quantifying the interference term

Area	Sub-area	References
Representation and Ranking	Projection Models	[van Rijsbergen 2004],[Melucci 2005a],[Melucci 2005b], [Piwowarski and Lalmas 2009a],[Piwowarski et al. 2008],[Piwowarski et al. 2010a],[Piwowarski et al. 2010c],[Piwowarski et al. 2010b], [Caputo et al. 2011], [Frommholz et al. 2010]
	Quantum Language Models	[Sordoni et al. 2013b], [Xie et al. 2015],[Zhang et al. 2018d] [Hou and Song 2009], [Hou et al. 2013], [Zhang et al. 2018b], [Zhang et al. 2018a], [Zhang et al. 2018d] , [Blacoe et al. 2013],[Blacoe 2015], [Basile and Tamburini 2017], [Li et al. 2018a]
	Quantum-inspired Ranking	[Zuccon et al. 2009], [Zuccon and Azzopardi 2010], [Zhao et al. 2011], [Zhang et al. 2011]
	Multimodal IR	[Wang et al. 2010a], [Kaliciak et al. 2011],[Kaliciak et al. 2013], [Gkoumas et al. 2018], [Zhang et al. 2018c], [Zhang et al. 2020]
	Quantum-inspired Representation Learning	[Aerts and Czachor 2004], [Bruza and Cole 2005], [Sordoni et al. 2013a], [Zuccon et al. 2011], [Wittek et al. 2014], [Li et al. 2018b], [Jaiswal et al. 2018], [Wang et al. 2019], [Li et al. 2019], [Buccio et al. 2018], [Tiwari and Melucci 2018]
User Interactions	Projection Models	[Melucci and White 2007], [Melucci 2008], [Piwowarski and Lalmas 2009a], [Frommholz et al. 2011], [Uprety et al. 2018], [Wang et al. 2018], [Wang et al. 2017]
	Feedback	[Li et al. 2015], [Zhang et al. 2016], [Li et al. 2016]
	Context Effects	[Wang et al. 2010b], [Wang et al. 2016], [Uprety and Song 2018], [Uprety et al. 2019b], [Wittek et al. 2016], [Uprety et al. 2019a], [Uprety et al. 2020]
Quantum-inspired Neural Networks		[Zhang et al. 2018a], [Li et al. 2018b], [Jaiswal et al. 2018], [Zhang et al. 2018d], [Zhang et al. 2019a] [Zhang et al. 2020], [Wang et al. 2019], [Li et al. 2019], [Zhang et al. 2019b], [Jiang et al. 2020]

Table 1. Review Summary

remains and document similarity approaches applied do not take the quantum advantage. One needs to devise a way to subscribe complex phases to documents and then calculate the interference terms.

The query fusion and query expansion approaches make use of superposition and interference phenomena, however it is difficult to get an intuitive explanation of how these two are coming into effect and providing the advantage over classical methods. The Contextual QLM (CQLM) and Adaptive-CQLM are promising applications of the QLM to incorporate user interactions, however they are outperformed by the state-of-art machine learning based methods.

The integration of the quantum framework to neural networks is promising and combines the representational complexity of neural networks with the probabilistic generalization provided by the quantum framework, especially when complex numbers are included. However, as we see in the results reported in this survey, the state-of-the-art neural networks outperform quantum-inspired models. A reason could be that the datasets used are mostly static, devoid of context, the human factor and its complexities, but it is not the case in real applications.

The cognitive experiments on order effects in document judgment provide a good insight into why quantum probability is useful in modeling human decision making. However, most of these experiments are only performed on small user collected samples and need to be conducted on real world search data. Also, they do not yet provide a way to make use of the order effect information to improve the effectiveness of IR systems.

We summarise the survey in Table 1 with the papers categorised into the sub-areas of IR mentioned in Figure 1. We also list papers which use quantum-inspired neural networks, encompassing all the other sub-areas.

5 FUTURE DIRECTIONS IN QUANTUM-INSPIRED IR

Quantum Theory was developed as a framework to explain the counter-intuitive behaviour of microscopic particles which could not be explained using traditional probability and logic models. Hence, the most important thing to consider

while applying the quantum framework to IR or any other computational sciences, is the existence of such non-classical data (which violates classical logic, e.g. Boolean logic). There is substantial evidence in behavioural sciences that data obtained from human information interaction and decision-making can be quantum-like data. Phenomena like conjunction and disjunction fallacy, violation of law of total probability (LTP), similarity effects [Tversky 1977], etc. can be investigated in user behaviour in IR. Conjunction and disjunction fallacies can exist in relevance judgment of documents. Although the QPRP incorporates the interference term in document ranking, it does not explicitly occur due to the violation of LTP. In fact, LTP violation can be investigated in IR when users make decisions under ambiguity and the cognitive state can be modelled as a superposition of different states.

Hence, the presence of quantum-like data and quantum phenomena in IR can completely change our understanding of the fundamentals of IR. A strong candidate to start with is to rethink the concept of relevance and the interpretation of probability of relevance. In the Cranfield methodology of building IR test collections, a common practice is to reject annotator disagreement as noise and fix one relevance label for a given document which has been judged by the majority of annotators. However, it is possible that both labels exist for the document for the same query, and the disagreement between annotators is due to their different contexts which leads to different semantic interpretations of the document. Or, there can be ambiguity in the document content and meta-data (e.g. arising out of its lack of credibility, novelty, etc.) which can put annotators in two minds, leading to different relevance labels. Hence a document needs to be modelled as both relevant and non-relevant for a query. We see that QT has tools to model two mutually exclusive states of a system as a single state (Superposition). Future work can begin with constructing such contextual datasets [Inel et al. 2014].

With such contextual data in hand, subspace generalization of vector spaces can be utilized to create more contextual and dynamic representations. Representing a document or a word with a subspace rather than a vector allows for representing its different aspects in the vectors which span the subspace. This can be instrumental in capturing meaning in the data in the same way as humans do (subject to multiple contexts, arising out of external factors or intrinsic cognitive biases). Integrating different Hilbert Spaces by using the tensor product is a useful technique to fuse different modes of information like text, images, audio, etc. as in multi-modal IR. It can also be useful in fusing different cognitive aspects of information (Polyrepresention) e.g., product reviews sentiment and brand credibility.

The presence of incompatibility in judgments can render many user models, which assume joint distributions between variables, as ineffective. For example, in [Lin and He 2009], a joint sentiment-topic model to detect sentiment and topics simultaneously from text is proposed. The joint probability of a word, topic and sentiment label assigned is written as $p(w, z, l) = p(w|z, l)p(z, l)$ where w is a word, z is a topic of the document and l is the sentiment label of the document. If there is incompatibility between the sentiment and topic of a document, then the predictions based on this model would not match the user's decisions [Bruza and Chang 2014]. This is something which can be captured using the quantum framework and quantum-inspired models will be better equipped to handle such cases.

This research can further benefit from a formal model of quantum-inspired neural networks with a theoretical basis and where the interference term and complex numbers naturally occur in the neural computations. We believe that best way forward in this field will be the integration of concepts and constructs from QT with the state-of-the-art machine learning models, e.g. neural networks. QT framework is best positioned to model human decisions under ambiguity and dynamic changes of context. Their fusion with QT can enhance their ability to model complex human behavioural data, building a platform for more human-centred Artificial Intelligence. It is known that the neural models suffer from the problem of generalisation. One way to tackle the problem is that multiple statistical hypotheses of a neural network can be preserved in the form of a superposition state in the middle layer and used by the model decision layer. Traditionally, a deep learning model is based on one hypothesis, thus limiting the model's generalisation ability. The output layer of

the deep learning model can integrate multiple statistical hypotheses that the hidden layer retained and learned during training. The nature of reserving multi-hypothesis at the same time is like quantum superposition. This makes for some exciting research prospects in the future.

The authors are hopeful that this literature survey is able to provide a clear picture of the quantum-inspired IR field and set a road-map for researchers to take this field forward.

REFERENCES

Diederik Aerts and Marek Czachor. 2004. Quantum aspects of semantic analysis and symbolic artificial intelligence. *Journal of Physics A: Mathematical and General* 37, 12 (2004), L123.

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* 16, 6 (2010), 345–379.

Ivano Basile and Fabio Tamburini. 2017. Towards Quantum Language Models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1840–1849.

William Blacoe. 2015. Semantic Composition Inspired by Quantum Measurement. In *Quantum Interaction*. Springer International Publishing, 41–53.

William Blacoe, Elham Kashefi, and Mirella Lapata. 2013. A Quantum-Theoretic Approach to Distributional Semantics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 847–857.

Niels Bohr. 1937. Causality and Complementarity. *Philosophy of Science* 4, 3 (1937), 289–298.

Pia Borlund. 2013. Interactive Information Retrieval: An Introduction. *Journal of Information Science Theory and Practice* 1 (2013).

Max Born. 1926. Quantum mechanics of collision processes. *Zeit fur Phys* 38 (1926), 803.

Peter Bruza and Vivien Chang. 2014. Perceptions of document relevance. *Frontiers in Psychology* 5 (2014), 612.

Peter Bruza and Richard J. Cole. 2005. Quantum Logic of Semantic Space: An Exploratory Investigation of Context Effects in Practical Reasoning. In *We Will Show Them! Essays in Honour of Dov Gabbay, Volume One*. 339–362.

Emanuele Di Buccio, Qiuchi Li, Massimo Melucci, and Prayag Tiwari. 2018. Binary Classification Model Inspired from Quantum Detection Theory. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval - ICTIR '18*. ACM Press.

Christopher J. C. Burges. 2010. From RankNet to LambdaRank to LambdaMART: An Overview.

Curt Burgess, Kay Livesay, and Kevin Lund. 1998. Explorations in context space: Words, sentences, discourse. *Discourse Processes* 25, 2-3 (1998), 211–257.

Jerome R. Busemeyer and Peter D. Bruza. 2012. *Quantum Models of Cognition and Decision* (1st ed.). Cambridge University Press.

Jerome R. Busemeyer, Emmanuel M. Pothos, Riccardo Franco, and Jennifer S. Trueblood. 2011. A quantum theoretical explanation for probability judgment errors. *Psychological Review* 118, 2 (2011), 193–218.

Annalina Caputo, Benjamin Piwowarski, and Mounia Lalmas. 2011. A Query Algebra for Quantum Information Retrieval. In *Proceedings of the 2nd Italian Information Retrieval (IIR) Workshop, Milan, Italy, January 27-28, 2011*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. 4171–4186.

Paul A. M. Dirac. 1982. *The Principles of Quantum Mechanics (International Series of Monographs on Physics)*. Clarendon Press.

Michael Eisenberg and Carol Barry. 1988. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science* 39, 5 (1988), 293–300.

Richard P. Feynman, Robert B. Leighton, and Matthew Sands. 2011. *The Feynman Lectures on Physics, Vol. III: The New Millennium Edition: Quantum Mechanics (Feynman Lectures on Physics (Paperback)) (Volume 3)*. Basic Books.

Ingo Frommholz, Birger Larsen, Benjamin Piwowarski, Mounia Lalmas, Peter Ingwersen, and Keith van Rijsbergen. 2010. Supporting Polyrepresentation in a Quantum-inspired Geometrical Retrieval Framework. In *Proceedings of the Third Symposium on Information Interaction in Context (IIiX '10)*. 115–124.

Ingo Frommholz, Benjamin Piwowarski, Mounia Lalmas, and Keith van Rijsbergen. 2011. Processing Queries in Session in a Quantum-Inspired IR Framework. In *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 751–754.

Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Jarvelin, Rosie Jones, YiquN Liu, Josiane Mothe, Wolfgang Nejdl, Isabella Peters, and Benno Stein. 2018. An Information Nutritional Label for Online Documents. *SIGIR Forum* 51, 3 (2018), 46–66.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2019. TANDA: Transfer and Adapt Pre-Trained Transformer Models for Answer Sentence Selection. *arXiv preprint arXiv:1911.04118* (2019).

Dimitris Gkoumas, Sagar Uprety, and Dawei Song. 2018. Investigating non-classical correlations between decision fused multi-modal documents. In *International Symposium on Quantum Interaction*. Springer, 163–176.

Andrew Gleason. 1957. Measures on the Closed Subspaces of a Hilbert Space. *Indiana University Mathematics Journal* 6, 4 (1957), 885–893.

Robert B. Griffiths. 2001. *Consistent Quantum Theory*. Cambridge University Press.

Carl W. Helstrom. 1969. Quantum detection and estimation theory. *Journal of Statistical Physics* 1, 2 (01 Jun 1969), 231–252.

Yuexian Hou and Dawei Song. 2009. Characterizing Pure High-Order Entanglements in Lexical Semantic Spaces via Information Geometry. In *QI*.

Yuexian Hou, Xiaozhao Zhao, Dawei Song, and Wenjie Li. 2013. Mining Pure High-order Word Associations via Information Geometry for Information Retrieval. *ACM Trans. Inf. Syst.* 31, 3, Article 12 (2013), 32 pages.

Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *Proceedings of International Semantic Web Conference*. Springer, 486–504.

Peter Ingwersen. 1996. Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory. *Journal of Documentation* 52, 1 (1996), 3–50.

Amit Kumar Jaiswal, Guilherme Holdack, Ingo Frommholz, and Haiming Liu. 2018. Quantum-like Generalization of Complex Word Embedding: A Lightweight Approach for Textual Classification. In *Proceedings of the Conference "Lernen, Wissen, Daten, Analysen", LWDA 2018, Mannheim, Germany, August 22-24, 2018*. 159–168.

Kalervo Järvelin and Peter Ingwersen. 2012. *User-oriented and cognitive models of information retrieval.* Taylor & Francis, United States, 47–64. Reprint of article in Encyclopedia of Library and Information Science, Taylor & Francis, 2010, 3. ed., p. 5521-5534.

Yongyu Jiang, Peng Zhang, Hui Gao, and Dawei Song. 2020. A Quantum Interference Inspired Neural Matching Model for Ad-hoc Retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '20)*.

Leszek Kaliciak, Dawei Song, Nirmalie Wiratunga, and Jeff Pan. 2013. Combining visual and textual systems within the context of user feedback. In *International Conference on Multimedia Modeling*. Springer, 445–455.

Leszek Kaliciak, Jun Wang, Dawei Song, Peng Zhang, and Yuexian Hou. 2011. Contextual image annotation via projection and quantum theory inspired measurement for integration of text and visual features. In *International Symposium on Quantum Interaction*. Springer, 217–222.

Jingfei Li, Peng Zhang, Dawei Song, and Yuexian Hou. 2016. An adaptive contextual quantum language model. *Physica A: Statistical Mechanics and its Applications* 456 (2016), 51–67.

Jingfei Li, Peng Zhang, Dawei Song, and Yue Wu. 2017. Understanding an enriched multidimensional user relevance model by analyzing query logs. *Journal of the Association for Information Science and Technology* 68, 12 (2017), 2743–2754.

Qiuchi Li, Jingfei Li, Peng Zhang, and Dawei Song. 2015. Modeling Multi-Query Retrieval Tasks Using Density Matrix Transformation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. 871–874.

Qiuchi Li, Massimo Melucci, and Prayag Tiwari. 2018a. Quantum Language Model-based Query Expansion. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '18)*. 183–186.

Qiuchi Li, Sagar Uprety, Benyou Wang, and Dawei Song. 2018b. Quantum-Inspired Complex Word Embedding. In *Proceedings of The Third Workshop on Representation Learning for NLP, Rep4NLP@ACL 2018, Melbourne, Australia, July 20, 2018*. 50–57.

Qiuchi Li, Benyou Wang, and Massimo Melucci. 2019. CNM: An Interpretable Complex-valued Network for Matching. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. 4139–4148.

Chenghua Lin and Yulan He. 2009. Joint Sentiment/Topic Model for Sentiment Analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*. 375–384.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers* 28, 2 (1996), 203–208.

Alexander I. Lvovsky. 2004. Iterative maximum-likelihood reconstruction in quantum homodyne tomography. *Journal of Optics B: Quantum and Semiclassical Optics* 6, 6 (2004), S556.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1101–1104.

Massimo Melucci. 2005a. Can vector space bases model context. *CEUR Workshop Proceedings* 151 (01 2005).

Massimo Melucci. 2005b. Context Modeling and Discovery Using Vector Space Bases. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05)*. 808–815.

Massimo Melucci. 2008. A basis for information retrieval in context. *ACM Transactions on Information Systems* 26, 3 (2008), 1–41.

Massimo Melucci. 2010. An Investigation of Quantum Interference in Information Retrieval. In *Advances in Multidisciplinary Retrieval*. Springer Berlin Heidelberg, 136–151.

Massimo Melucci and Ryen W. White. 2007. Utilizing a geometry of context for enhanced implicit feedback. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07*. ACM Press.

Donald Metzler and W Bruce Croft. 2005. A Markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. 472–479.

Bhaskar Mitra and Nick Craswell. 2018. An Introduction to Neural Information Retrieval. *Foundations and Trends® in Information Retrieval* (2018), 1–117.

Kezban Dilek Onal, Ye Zhang, Ismail Sengor Altingovde, Md Mustafizur Rahman, Pinar Karagoz, Alex Braylan, Brandon Dang, Heng-Lu Chang, Henna Kim, Quinten McNamara, Aaron Angert, Edward Banner, Vivek Khetan, Tyler McDonnell, An Thanh Nguyen, Dan Xu, Byron C. Wallace, Maarten de Rijke, and Matthew Lease. 2018. Neural information retrieval: at the end of the early years. *Information Retrieval Journal* 21, 2 (01 Jun 2018), 111–182.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

Itamar Pitowsky. 2006. Quantum mechanics as a theory of probability. In *Physical theory and its interpretation*. Springer, 213–240.

Benjamin Piwowarski, Ingo Frommholz, Mounia Lalmas, and Keith van Rijsbergen. 2010a. What Can Quantum Theory Bring to Information Retrieval. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*. 59–68.

Benjamin Piwowarski, Ingo Frommholz, Yashar Moshfeghi, Mounia Lalmas, and Keith van Rijsbergen. 2010b. Filtering Documents with Subspaces. In *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 615–618.

Benjamin Piwowarski and Mounia Lalmas. 2009a. A Quantum-Based Model for Interactive Information Retrieval. In *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 224–231.

Benjamin Piwowarski and Mounia Lalmas. 2009b. Structured Information Retrieval and Quantum Theory. In *Proceedings of the 3rd International Symposium on Quantum Interaction (QI '09)*. Springer-Verlag, Berlin, Heidelberg, 289–298.

Benjamin Piwowarski, Mounia Lalmas, Ingo Frommholz, and Keith van Rijsbergen. 2010c. Exploring a Multidimensional Representation of Documents and Queries. In *Proceedings of Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO '10)*. Paris, France, France, 57–60.

Benjamin Piwowarski, Andrew Trotman, and Mounia Lalmas. 2008. Sound and Complete Relevance Assessment for XML Retrieval. *ACM Trans. Inf. Syst.* 27, 1, Article 1 (2008), 37 pages.

Emmanuel Pothos and Jerome Busemeyer. 2011. A quantum probability explanation for violations of symmetry in similarity judgments. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 33.

Emmanuel M. Pothos and Jerome R. Busemeyer. 2009. A quantum probability explanation for violations of 'rational' decision theory. *Proceedings of the Royal Society B: Biological Sciences* 276, 1665 (2009), 2171–2178.

Emmanuel M. Pothos, Jerome R. Busemeyer, and Jennifer S. Trueblood. 2013. A quantum geometric model of similarity. *Psychological Review* 120, 3 (2013), 679–696.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

Stephen E. Robertson. 1977. The Probability Ranking Principle in IR. *Journal of Documentation* 33, 4 (1977), 294–304.

Ian Ruthven. 2009. Interactive information retrieval. *Annual Review of Information Science and Technology* 42, 1 (2009), 43–91.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 440–450.

Marc Sloan and Jun Wang. 2015. Dynamic information retrieval: Theoretical framework and application. In *Proceedings of the 2015 International Conference on the theory of Information Retrieval*. 61–70.

Dawei Song, Mounia Lalmas, C. J. van Rijsbergen, Ingo Frommholz, Benjamin Piwowarski, Jun Wang, Peng Zhang, Guido Zuccon, Peter Bruza, S M Yasir Arafat, Leif Azzopardi, Emanuele Di Buccio, Alvaro Francisco Huertas-Rosero, Yuexian Hou, Massimo Melucci, and Stefan M. Rüger. 2010. How Quantum Theory Is Developing the Field of Information Retrieval. In *AAAI Fall Symposium: Quantum Informatics for Cognitive, Social, and Semantic Processes*.

Alessandro Sordoni, Jing He, and Jian-Yun Nie. 2013a. Modeling Latent Topic Interactions Using Quantum Interference for Information Retrieval. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management (CIKM '13)*. 1197–1200.

Alessandro Sordoni, Jian-Yun Nie, and Yoshua Bengio. 2013b. Modeling Term Dependencies with Quantum Language Models for IR. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. 653–662.

Alistair Sutcliffe and Mark Ennis. 1998. Towards a cognitive theory of information retrieval. *Interacting with Computers* 10, 3 (1998), 321–351.

Prayag Tiwari and Massimo Melucci. 2018. Towards a Quantum-Inspired Framework for Binary Classification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. 1815–1818.

Jennifer S. Trueblood and Jerome R. Busemeyer. 2011. A Quantum Probability Account of Order Effects in Inference. *Cognitive Science* 35, 8 (2011), 1518–1552.

Amos Tversky. 1977. Features of similarity. *Psychological Review* 84, 4 (1977), 327–352.

Hisaharu Umegaki. 1962. Conditional expectation in an operator algebra. IV. Entropy and information. *Kodai Math. Sem. Rep.* 14, 2 (1962), 59–85.

Sagar Uprety, Shahram Dehdashti, Lauren Fell, Peter Bruza, and Dawei Song. 2019a. Modelling dynamic interactions between relevance dimensions. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. 35–42.

Sagar Uprety, Dimitris Gkoumas, and Dawei Song. 2019b. Investigating Bell Inequalities for Multidimensional Relevance Judgments in Information Retrieval. In *Quantum Interaction*. Springer International Publishing, Cham, 177–188.

Sagar Uprety and Dawei Song. 2018. Investigating order effects in multidimensional relevance judgment using query logs. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*. 191–194.

Sagar Uprety, Yi Su, Dawei Song, and Jingfei Li. 2018. Modeling Multidimensional User Relevance in IR Using Vector Spaces. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. 993–996.

Sagar Uprety, Prayag Tiwari, Shahram Dehdashti, Lauren Fell, Dawei Song, Peter Bruza, and Massimo Melucci. 2020. Quantum-Like Structure in Multidimensional Relevance Judgements. In *Advances in Information Retrieval*. Springer International Publishing, Cham, 728–742.

Cornelis Joost van Rijsbergen. 2004. *The geometry of information retrieval*. Cambridge University Press.

John von Neumann. 1955. *Mathematical Foundations of Quantum Mechanics*. Princeton University Press.

Benyou Wang, Qiuchi Li, Massimo Melucci, and Dawei Song. 2019. Semantic Hilbert space for text representation learning. In *Proceedings of The World Wide Web Conference*. 3293–3299.

Benyou Wang, Peng Zhang, Jingfei Li, Dawei Song, Yuexian Hou, and Zhenguo Shang. 2016. Exploration of Quantum Interference in Document Relevance Judgement Discrepancy. *Entropy* 18, 12 (Apr 2016), 144.

Jun Wang, Dawei Song, and Leszek Kaliciak. 2010a. Tensor Product of Correlated Textual and Visual Features: A Quantum Theory Inspired Image Retrieval Framework. In *AAAI-Fall 2010 Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes (QI 2010)*. 109–116.

Jun Wang, Dawei Song, Peng Zhang, Yuexian Hou, and Peter Bruza. 2010b. Explanation of relevance judgement discrepancy with quantum interference. In *AAAI-Fall 2010 Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes (QI 2010)*. 117–124.

Jun Wang and Jianhan Zhu. 2009. Portfolio theory of information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 115–122.

Panpan Wang, Yuexian Hou, Jingfei Li, Yazhou Zhang, Dawei Song, and Wenjie Li. 2017. A quasi-current representation for information needs inspired by Two-State Vector Formalism. *Physica A: Statistical Mechanics and its Applications* 482 (2017), 627–637.

Panpan Wang, Tianshu Wang, Yuexian Hou, and Dawei Song. 2018. Modeling Relevance Judgement Inspired by Quantum Weak Measurement. In *Advances in Information Retrieval*. Springer International Publishing, Cham, 424–436.

Zheng Wang and Jerome R Busemeyer. 2013. A quantum question order model supported by empirical tests of an a priori and precise prediction. *Topics in Cognitive Science* 5, 4 (2013), 689–710.

Peter Wittek, Bevan Koopman, Guido Zuccon, and Sándor Darányi. 2014. Combining Word Semantics within Complex Hilbert Space for Information Retrieval. In *Quantum Interaction*. Springer Berlin Heidelberg, Berlin, Heidelberg, 160–171.

Peter Wittek, Ying-Hsang Liu, Sándor Darányi, Tom Gedeon, and Ik Soo Lim. 2016. Risk and Ambiguity in Information Seeking: Eye Gaze Patterns Reveal Contextual Behavior in Dealing with Uncertainty. *Frontiers in Psychology* 7 (2016).

Mengjiao Xie, Yuexian Hou, Peng Zhang, Jingfei Li, Wenjie Li, and Dawei Song. 2015. Modeling Quantum Entanglements in Quantum Language Models. In *Proceedings of Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI'15)*. 1362–1368.

Grace Hui Yang, Marc Sloan, and Jun Wang. 2016b. *Dynamic Information Retrieval Modeling (Synthesis Lectures on Information Concepts, Retrieval, and S)*. Morgan & Claypool Publishers.

Liu Yang, Qingyao Ai, Jiafeng Guo, and W Bruce Croft. 2016a. aNMM: Ranking short answer texts with attention-based neural matching model. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. 287–296.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*. 5754–5764.

Lipeng Zhang, Peng Zhang, Xindian Ma, Shuqin Gu, Zhan Su, and Dawei Song. 2019b. A generalized language model in tensor space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7450–7458.

Peng Zhang, Jingfei Li, Benyou Wang, Xiaozhao Zhao, Dawei Song, Yuexian Hou, and Massimo Melucci. 2016. A Quantum Query Expansion Approach for Session Search. *Entropy* 18 (2016), 146.

Peng Zhang, Jiabin Niu, Zhan Su, Benyou Wang, Liqun Ma, and Dawei Song. 2018a. End-to-end quantum-like language models with application to question answering. In *Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence (AAAI'18)*. 5666–5673.

Peng Zhang, Dawei Song, Xiaozhao Zhao, and Yuexian Hou. 2011. Investigating query-drift problem from a novel perspective of photon polarization. In *Conference on the Theory of Information Retrieval*. Springer, 332–336.

Peng Zhang, Zhan Su, Lipeng Zhang, Benyou Wang, and Dawei Song. 2018d. A Quantum Many-body Wave Function Inspired Language Modeling Approach. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. 1303–1312.

Yazhou Zhang, Qiuchi Li, Dawei Song, Peng Zhang, and Panpan Wang. 2019a. Quantum-Inspired Interactive Networks for Conversational Sentiment Analysis. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*. 5436–5442.

Yazhou Zhang, Dawei Song, Xiang Li, and Peng Zhang. 2018b. Unsupervised Sentiment Analysis of Twitter Posts Using Density Matrix Representation. In *Advances in Information Retrieval*. Springer International Publishing, Cham, 316–329.

Yazhou Zhang, Dawei Song, Xiang Li, Peng Zhang, Panpan Wang, Lu Rong, Guangliang Yu, and Bo Wang. 2020. A Quantum-like Multimodal Network Framework for Modeling Interaction Dynamics in Multiparty Conversational Sentiment Analysis. *Information Fusion* (2020).

Yazhou Zhang, Dawei Song, Peng Zhang, Panpan Wang, Jingfei Li, Xiang Li, and Benyou Wang. 2018c. A quantum-inspired multimodal sentiment analysis framework. *Theoretical Computer Science* 752 (2018), 21–40.

Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. 2014. Multidimensional Relevance Modeling via Psychometrics and Crowdsourcing. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. 435–444.

Xiaozhao Zhao, Peng Zhang, Dawei Song, and Yuexian Hou. 2011. A Novel Re-ranking Approach Inspired by Quantum Measurement. In *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 721–724.

Liron Zighelnic and Oren Kurland. 2008. Query-drift Prevention for Robust Query Expansion. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. 825–826.

Guido Zuccon and Leif Azzopardi. 2010. Using the Quantum Probability Ranking Principle to Rank Interdependent Documents. In *Proceedings of the 32nd European Conference on Advances in Information Retrieval (ECIR'2010)*. Springer-Verlag, Berlin, Heidelberg, 357–369.

Guido Zuccon, Leif A. Azzopardi, and Keith van Rijsbergen. 2009. The Quantum Probability Ranking Principle for Information Retrieval. In *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 232–240.

Guido Zuccon, Benjamin Piwowarski, and Leif Azzopardi. 2011. On the use of Complex Numbers in Quantum Models for Information Retrieval. In *Advances in Information Retrieval Theory*. Springer Berlin Heidelberg, Berlin, Heidelberg, 346–350.